Structural Study of the Adenylation Domain by Molecular Dynamics Simulation

Joanne Frances Thalassinou BSc (Hons), MRes

A thesis submitted for the degree of Doctor of Philosophy

University of Warwick

Department of Chemistry

February 2012

Contents

List of Figures

List of Tables

Declaration

I hereby declare that this thesis, submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy and entitled ”Structural Study of the Adenylation Domain by Molecular Dynamics Simulation”, represents my own work and has not been previously submitted to any other institution for any degree, diploma or other qualification.

Joanne Frances Thalassinou
February 2012

Acknowledgements

I would like to thank my supervisor Professor Mark Rodger for his support and guidance. This work would not have been possible without the love and support of my family and friends. In particular I would like to thank my parents, Andrea, Lee, Ethan and Vib.

Most importantly I would like to acknowledge my husband, Kostas, for his love, support and patience.

Joanne Frances Thalassinou

Summary

As antibiotic resistance is increasing more rapidly than new antibiotics are produced and/or discovered, there is an increasing need to identify new ways to design novel antibiotics. A potential avenue for this, is the exploitation of Nonribosomal Peptide Synthetases (NRPSs) from bacteria and fungi which biosynthesise structurally complex biologically active peptide products, including numerous potential antibiotics and other molecules with pharmacologically attractive properties. In order to do so, however, a detailed molecular understanding of NRPSs is required.

NRPSs are modular proteins, with each module comprising domains that each perform specific functions to select, activate, alter (optional) and combine amino/hydroxyl acid substrates to form a specific peptide product. The Adenylation domain (A domain) specifically selects and activates the substrate through a two step reaction. In the first half reaction, a highly reactive aminoacyl adenylate is formed by reaction with Mg-adenosine triphosphate (ATP) resulting in the release of pyrophosphate. In the second half reaction the A domain binds the phosphopantetheinyl (PPant) arm of the downstream domain, the Peptidyl Carrier Protein (PCP) domain. The terminal thiol of the PPant arm attacks the activated aminoacyl group displacing adenosine monophosphate (AMP), leaving the amino acid substrate tethered to the PCP domain as a thioester.

The A domain is of particular interest as a target for engineering approaches as it is considered to be the primary determinant of substrate specificity. Little is understood, however, about the molecular basis of substrate selectivity or how the dynamics of the domain enable the two part reactions to take place.

In 1997, the first A domain structure was determined; the L-phenylalanine (L-Phe) activating A domain (PheA) of the Gramicidin S synthetase from *Bacillus brevis*. All of the A domain structures determined to date are either unligated (apo form) or co-crystallised with reactants or products from the first half reaction. The NRPS A domains are members of the adenylate-forming superfamily which have been structurally characterised in three states, apo, with the first half reaction and second half reaction ligands. Comparison between these structures, suggested these enzymes use a domain alternation strategy to reconfigure a single active site to perform two different reactions. While the A domains have only been determined in the adenylate-forming conformation, similarities between members of the adenylate-forming superfamily suggest NRPS A domains may exploit of a similar strategy of domain alternation to reconfigure the enzyme's single active site.

To date, no molecular simulation study of any NRPS A domain has been reported in the literature. In this study, molecular dynamics (MD) simulations of the PheA have been carried out in the apo form, with the cognate substrate, and with noncognate substrates, to understand the molecular basis of substrate specificity and the effect of the substrate on the dynamics of the protein.

Inter-domain rotation was observed in the apo and cognate holo simulations and with one of the noncognate substrates, L-Thr. This motion occurred between the A_{core} domain and A_{sub} domain or part of the A_{sub} domain. The rotation observed in the simulations with the cognate substrate creates a widening between the two domains of PheA on the side of the enzyme where the PPant arm is thought to bind. Results from one of the cognate holo simulations suggests the A3 motif loop may be important in stabilising the A domain to increase the domain rotation or maintaining the opening through with PPant is proposed to access the active site.

Results from one of the noncognate substrate simulations, L-Asp substrate, suggests a role for the A3 motif loop in the removal of noncognate ligands from the binding site. Results from the simulation with noncognate substrate L-Tyr also suggest that interaction of the substrate with the key Asp and Lys binding pocket residues may be required for rotation of the A_{sub} domain can occur.

A homology model of the second A domain of the NRPS that forms Coelichelin has built and it is shown that the core regions of the model are stable in the MD simulations carried out in the apo form, with the cognate ligand (L-Thr) and noncognate ligands (L-Ser and L-Val). Some domain rotation was observed in the simulations with L-Thr and L-Ser. The findings from this study support the suggestion that interaction between the key Asp and Lys binding pocket residues and the substrate may be required for domain rotation.

This work presented in this thesis useful insight into the dynamics of the A domain and provides evidence for the role of the conserved A3 motif loop in both domain rotation and removal of noncognate ligands from the binding pocket.

Abbreviations

A

A Adenylation.

AMP Adenosine Monophosphate.

atm Atmospheres.

ATP Adenosine Triphosphate.

B

BLOSUM BLOcks SUubstitution Matrix.

BPTI Bovine Pancreatic Trypsin Inhibitor.

C

C Condensation.

C-Mt C-Methylation.

CG Conjugate Gradient.

Cy Heterocyclisation.

D

DOPE Discrete Optimized Protein Energy.

E

E Epimerisation.

F

F Formylation.

FAD flavin adenine dinucleotide.

fs Femtoseconds.

G

GA genetic algorithm.

GOR Garnier, Osgusthorpe and Robson.

GROMACS Groningen Machine for Chemical Simualtions.

H

HMM Hidden Markov Model.

K

K Kelvin.

L

LGA Lamarckian genetic algorithm.

LS local search.

M

MC Monte Carlo.

MD Molecular Dynamics.

MSM Markovian state models.

N

N-Mt N-Methylation.

nm nanometre.

NOE nuclear Overhauser effect.

NRP Nonribosomal peptide.

NRPS Nonribosomal peptide synthetase.

O

Ox Oxidation.

P

PAM Percentage Accepted Mutation.

PBC Periodic Boundary Conditions.

PCA principal components analysis.

PCP Peptidyl Carrier Protein.

PDB Protein Data Bank.

PDF Probability Density Function.

PME particle-mesh Ewald.

Ppant 4'-phosphopantetheinyl.

ps Picoseconds.

PSSM Position Specific Score Matrix.

R

Red Reduction.

RMSD Root Mean Square Deviation.

RMSF root mean square fluctuations.

S

SAM S-adenosyl methionine.

SD Steepest Descent.

SPC Simple Point Charge.

STMV satellite tobacco mosaic virus.

T

T Thiolation domain.

Te Thioesterase.

V

vdW van der Waals.

VMD Visual Molecular Dynamics.

Chapter 1

Introduction

Small peptide natural products have a range of powerful biological activities and are critical elements of modern therapy. Mainly synthesised by microorganisms they are produced either by the ribosomal machinery or by gigantic multi-domain enzymes called nonribosomal peptide synthetases (NRPS) using a thiotemplated mechanism. Nonribosomal peptide (NRP) natural products are structurally diverse secondary metabolites thought to be produced primarily to offer the host organism a survival advantage and which have been optimised to perform a certain function(s) over years of evolution. The diverse sphere of action they possess includes antibiotic, antifungal, immunosuppressive and cytostatic activity[2]. NRPs, especially those with antibiotic activity, have been and continue to be of tremendous pharmacological importance either as therapeutic agents or as promising scaffolds for the development of substances with novel activities. NRPSs are composed of catalytic domains arranged into modules. Each module is responsible for the specific incorporation of a proteinogenic or non-proteinogenic amino acid monomer into the peptide product. This relatively simple biosynthetic logic generates peptides of high structural complexity[3]. The modular multi-domain architecture of these synthetases makes them amenable to genetic manipulation and is one strategy for the production of "novel natural products". Many of the key principles of nonribosomal peptide synthesis have been determined using biochemical and genetic studies. A comprehensive understanding of the molecular basis of the numerous protein-protein recognition events underpinning the mechanism of nonribosomal peptide synthesis is necessary to realise the potential of producing novel products with predefined characteristics by genetic modification of the synthetases[4].

1.1 Nonribosomal Peptides

Nonribosomal peptides (NRP) are structurally diverse complex peptide secondary metabolites of low molecular weight (see figure 1.1 for examples). The sheer bulk of the multi-domain NRPSs and the rate at which the products are synthesised places a limitation on the size of the peptide produced[7]. NRPs display useful therapeutic and agriculturally important activities including antibiotic, antifungal, cytostatic, immunosuppressive, iron chelating, pigment producing and toxic properties. Nonribosomally synthesised antibiotics include:

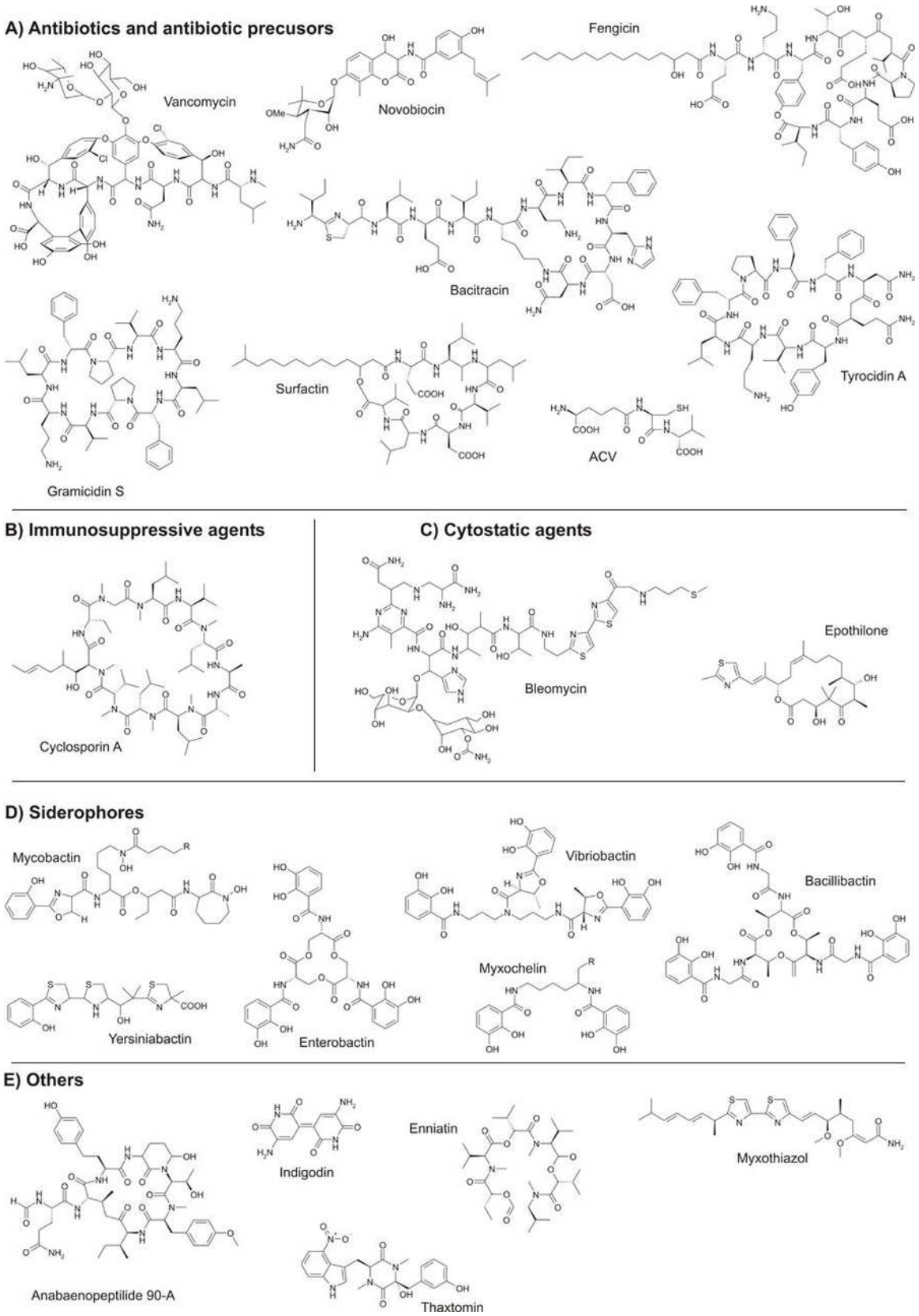


Figure 1.1: **Examples of nonribosomally synthesised peptides classified into five groups (A-E) according to their biological properties.** **A**: antibiotics and antibiotic precursors; **B**: immunosupressive agents; **C**: cystostatic (anti-cancer) agents; **D**: siderophores; and **E**: peptides with other properties, which include phytotoxin (enniatin) and pigment producing (indigodin) properties. Image adapted from scheme 1 in[5] and figure 1 in[6].

gramicidin S[8–10], vancomycin[11], daptomycin[12–14], and the penicillin and cephalosporin tripeptide precursor ACV[15]. Cyclosporin A produced by *Tolypocladium niveum* is a NRP with immunosuppressive activity routinely used in transplant aftercare[16]. Thiocoraline[17] has potent antitumor activity and is currently undergoing clinical trials for use in cancer therapy[18]. Yersiniabactin[19], vibriobactin[20] and enterobactin[21] are all siderophores, iron scavengers that are produced under iron limiting conditions. This chelation of iron by bacteria is vital for their survival and is often a virulence determinant in pathogens[21].

The great structural diversity exhibited by NRPs distinguishes them from peptides synthesised ribosomally and can in part, yet not exclusively, be attributed to the array of precursors NRPSs can utilise. Unlike ribosomal protein synthesis which is limited to the 22 proteinogenic α-amino acid building blocks, NRPSs have been shown to incorporate several hundred substrates[22], including many non-proteinogenic amino, aryl carboxylic and α-hydroxy acids. These include L-hf ornithine found in coelichelin[23], dihydroxyphenylglycine (DHPG) in vancomycin, (*4R*)-4-[(*E*)-2-butenyl]-4-methyl-L-threonine (Bmt) in cyclosporin A and 2,3-dihydroxybenzoate (DHB) in vibriobactin. The structures can be linear (myxothiazol), macrocyclic (tyrocidin A[24]), branched macrocyclic (fengycin), or dimers (gramicidin S) or trimers (enterobactin) of identical structural elements. NRPs often contain small heterocyclic rings such as thiazole (epothilone) and oxazoline (vibriobactin), and may contain *N*-formylations (anabaenopeptilide 90-A[25]), *N*-methylations (cyclosporin A), acylations and glycosylations (vancomycin). The majority of known NRPs contain either unusual or modified amino acids either at their N- or C-termini, which suggests preselectivity of these compounds for stability and biological activity[5]. The vast structural diversity of these natural products is strictly associated with their biological function.

1.2 Nonribosomal Peptide Synthetases

In ribosomal protein synthesis, protein structure is determined by the genetic code. In contrast, the multi-domain proteins called nonribosomal peptide synthetases (NRPS) act both as the structural template and the assembly line machinery in nonribosomal synthesis.

Found in bacteria and fungi, NRPSs assemble peptides by the repetitive condensation of simple monomers using a strategy termed the multiple thiotemplated mechanism[26–28].

The complete synthesis of a NRP can be performed either by a single synthetase, as is usually the case in fungi, or by a series of structurally distinct synthetases, as is often seen in bacteria, the encoding genes of which are almost always organised in an operon. NRPSs are organised into modules, each one responsible for the specific recognition, activation and incorporation of one monomer into the nascent peptide chain. The number and order of modules in the synthetase usually dictates the primary sequence of the peptide product; these NRPSs are referred to as type A. NRPSs that do not follow this linear logic however, cannot be considered as rare exceptions but rather as variations of the common NRPS repertoire designed to increase the biosynthetic potential of the synthetases[5]. The modules of an iterative NRPS (type B) are used in a sequential repeated manner to produce peptides containing multiple copies of identical structural elements. Iterative NRPSs include those that synthesise gramicidin S, a pentapeptide dimer, and enterobactin, a dipeptide trimer. Nonlinear NRPSs can use one or more of their modules more than once to generate peptides that contain repeats of specific precursors. Iterative NRPSs commonly contain unusual arrangements of the NRPS domains. The first modules of nonlinear synthetases syringomycin and coelichelin of *Stigmatella aurantiaca* and *Streptomyces coelicolor* is used twice to produce the tripeptide and tetrapeptide products, respectively[5].

NRPS modules are composed of individual domains with defined functions that, when present on the same synthetase, are separated by short spacer regions of about 15 amino acids. These regions are homologous to classical protein linkers, see figure 1.2 for the location and structure of these linker regions in the surfactin synthetase SrfA-C termination module. These physically linked NRPS domains retain their functionality when excised and expressed heterologously as separate units. NRPSs that synthesise siderophores commonly have fewer physical linkages between the domains. Such domains are referred to as stand-alone or free-standing. The interactions between physically linked domains are termed intramolecular or *in cis* interactions, and those between distinct stand-alone domains, or domains on different synthetases, as intermolecular or *in trans* interactions[29].

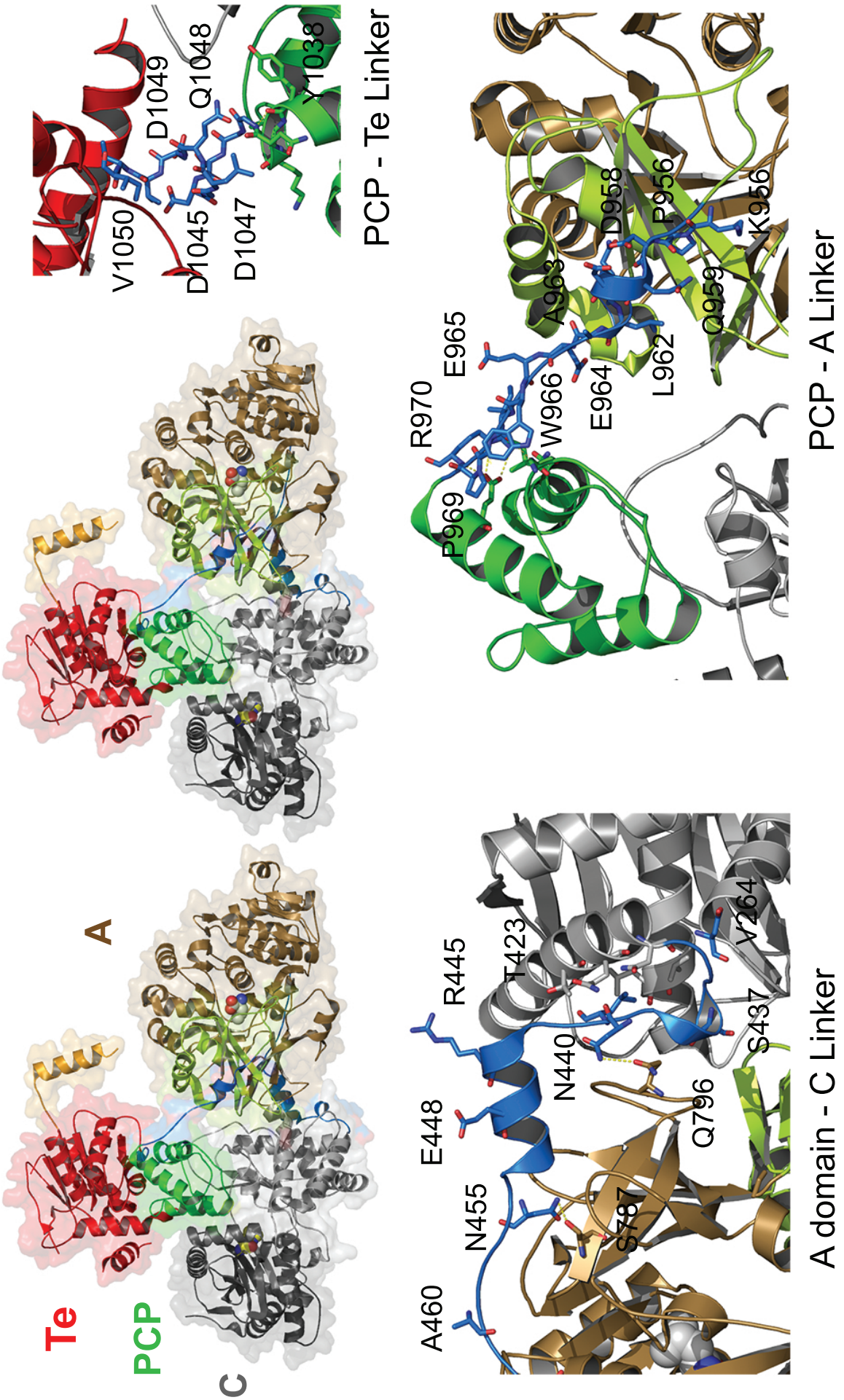


Figure 1.2: **The surfactin synthetase SrfA-C termination module.** The overall orientation of the domains in the termination module is shown. The C domain (grey), the A domain (brown, Acore, yellow Asub), the PCP domain (green), the Te domain (red), the C-terminal peptide tag (orange) and the linkers (blue). The C-A linker (shown above in the A-C orientation) consists of 32 residues ,eleven of which form an α-helical segment, connecting the C and A domains. The A-PCP linker (shown above in the PCP-A orientation) comprises 15 resides and links the A_{sub} domain of the A domain to the PCP domain. Nine residues form the PCP-Te linker joining the PCP and Te domains. Secondary structure elements and the solvent accessible surface area displayed. The substrates are displayed as CPK to highlight the active sites of the A domain and C domain.

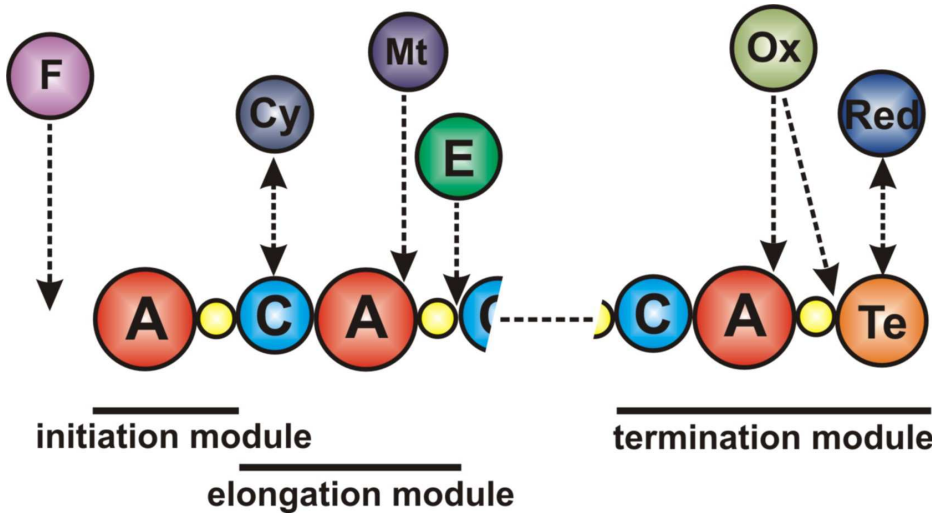


Figure 1.3: **The ten NRPS domains.** Where; F - formylation, A - adenylation, C - condensation, Cy - cyclisation, E - epimerisation, Ox - oxidation, Te - thioesterase, and Red - Reductase. PCP domains are coloured yellow. Mt - represents both N- and C-Methylation domains. Image adapted from figure 1 in reference [30].

A typical round of peptide initiation and elongation is illustrated in figure 1.4 and proceeds as follows. The Adenylation (A) domain of module 1, A_1, specifically selects and activates the amino acid substrate, forming a highly reactive aminoacyl adenylate by reaction with Mg-Adenosine triphosphate (ATP). The A domain is therefore considered to be the primary determinant of substrate specificity. After pyrophosphate (PP_i) is released the A domain binds the phosphopantetheinyl (Ppant) arm of the downstream Peptidyl Carrier Protein (PCP) domain [1], PCP_1. The terminal thiol of this arm attacks the activated aminoacyl group displacing adenosine monophosphate (AMP), leaving the amino acid substrate tethered to the PCP domain as a thioester. The downstream Condensation (C) domain of the adjacent module, module 2, catalyses peptide bond formation between substrates bound to the PCPs of modules 1 and 2. The C domain catalyses the nucleophilic attack of the substrate bound to the PCP_2 domain on the activated thioester of the substrate bound on the upstream PCP_1 domain[31]. This condensation reaction results in the covalent attachment of the peptidyl product to the PCP_2 domain and the release of the sulfhydryl group of the PPant moiety of the PCP_1 domain. The C_3 domain then forms a peptide bond between the peptidyl product from the last reaction, now tethered to PCP_2, and the substrate attached to the PCP_3 module. The tripeptide produced from this reaction is now tethered to the PCP_3 domain. The peptidyl chain continues to grow in this fashion until all of the substrates

[1] the PCP domain is also referred to as the thiolation (T) domain

are incorporated, at which point the chain is released from the last domain in the final (or termination) module either by hydrolysis or by cyclisation. This is usually achieved by a thioesterase (Te) domain in a two stage reaction. An acyl-*O*-TE-enzyme intermediate is formed, then subsequently attacked either by water[32] or a peptide-internal nucleophile[33]. This produces either a linear or a macrocyclic peptide. The favoured mechanism appears to be macrocyclic release.

A schematic of the modular organisation of the ten known NRPS domains is shown in figure 1.3. Of these ten domains, only three - the A, PCP and C domains - are needed to perform all the functions required for one complete cycle of elongation. These core domains are arranged in the order C-A-PCP in a minimal elongation module. A minimal initiation, or starter module, consists of A-PCP, and a standard termination module consists of an elongation module followed by a Te domain, C-A-PCP-Te. Optional NRPS domains include the Epimerisation (E), Heterocyclisation (Cy), N- and C-Methylation (N-Mt and C-Mt), Oxidation (Ox), N-formyltetrahydrofolate-dependent formyltransferase (F) and Reduction (Red) domains. While the majority are optional editing domains that can be present in any elongation module, the Red domain can only be located in a termination module as a replacement of a Te domain. All of the NRPS domains can be identified from primary sequence data by the location of a series of highly conserved motifs (shown in table 7.1 in appendix 7.1.2). A complete summary of the reactions catalysed by each of the NRPS domains, including the optional domains, is shown in figure 1.5.

The speed, order and uni-directionality of the peptide elongation reaction are controlled by the C domain[34]. This domain possesses a donor site for the electrophile (the substrate from the upstream PCP domain) and an acceptor site for the nucleophile (the substrate on the downstream PCP domain). Strong stereoselectivity[31,34,35] and a degree of selectivity towards the side chain of the aminoacyl thioester[35] are observed at the C domain acceptor site. The donor site exhibits broader substrate specificity[31,36].

Structural models have been determined for each of the main NRPS domains (A, PCP, C and Te), for a PCP-C didomain[37]. The structure of the entire NRPS termination module (C-A-PCP-Te) from surfactin synthetase SrfA-C was determined[38] in 2008, see figure 1.2.

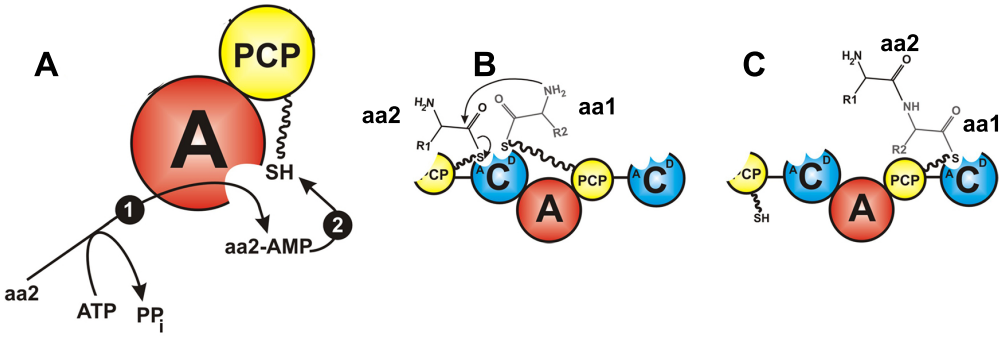


Figure 1.4: **The reaction sequence of peptide chain elongation.** The Ppant prosthetic group is represented by the zigzagged line. **A**1 The A domain selects a specific substrate and catalyses formation of the amino acyl adenylate by ATP hydrolysis. **A**2 The acyl moiety is transferred to the thiol group of the PCP PPant prosthetic group. **B** Transfer of the substrate to the acceptor site of the upstream C domain is facilitated by movement of the acyl-S-Ppant. Peptide bond formation with the amino acyl or peptidyl chain of the preceding PCP domain is then catalysed. **C** The donor site of the downstream C domain (from the subsequent module) is where the elongation cycle is completed. Image adapted from[5].

1.2.1 Polyketide Synthetases

The biosynthetic strategy of the modular architecture of NRPSs is comparable to that of the polyketide synthetases (PKS) of secondary metabolism and fatty acid synthetases (FAS) of primary metabolism. These similarities facilitate the transition between NRPS and PKS modules in NRPS-PKS hybrid megaenzymes[39]. The PKS modules integrate acetate and propionate into the peptide chain. Numerous hybrid NRPS-PKS systems have been discovered and the ratio of NRPS to PKS modules can vary greatly. These hybrid systems produce structures with biological activities similar to NRPs. NRPS-PKS products include the anticancer molecules bleomycin A2[40] and epothilone[41,42].

1.3 Molecular Engineering Approaches

Almost all peptide-based antibiotics are made by NRPSs[43]. The growing number of pathogenic bacterial strains resistant to antibiotics, especially in hospitals, is of great medical, societal and governmental concern. Historically the introduction and use of newly developed effective and safe antibiotics is followed by the rapid development of resistance mechanisms in the target organism[44]. Unfortunately, antibiotic use selects for bacterial strains with de-

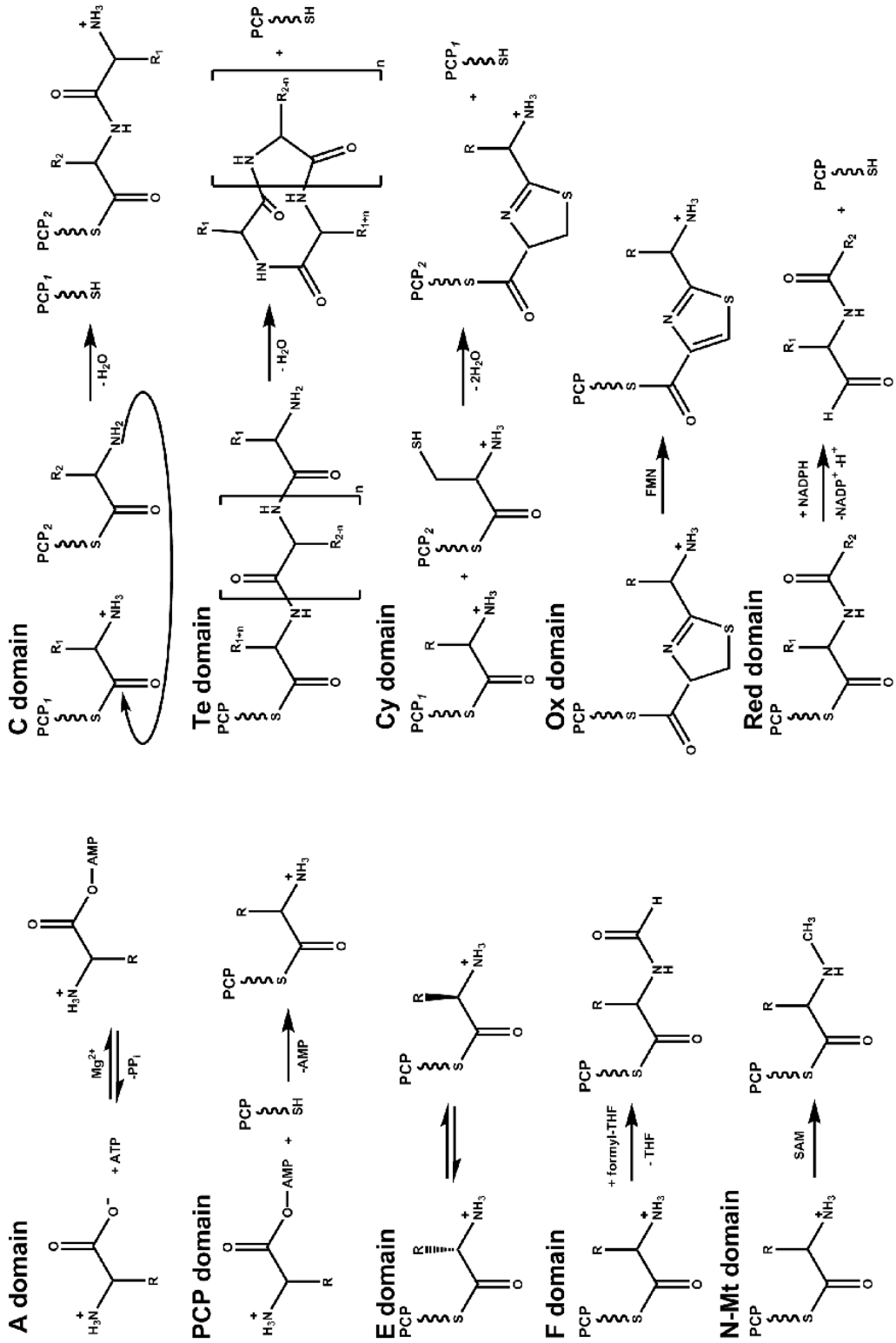


Figure 1.5: **Reactions catalysed by the NRPS domains.** Image adapted from figure 2 of[6].

veloped resistance mechanisms and the way in which bacteria transfer genetic information between each other has lead to the formation of multi-drug resistant pathogenic bacterial strains. Thus the average pre-resistance life span of any antibiotic is short and the development and discovery of new antibiotics a constant requirement.

The multidomain, modular, assembly line architecture of NRPSs makes them amenable to reprogramming in order to synthesise novel antibiotic products. Strategies to generate designer products include the formation of hybrid NRPSs by specifically recombining the domains and modules, or altering domain specificity. Hybrid NRPSs have been formed by: fusing domains together, e.g. A-PCP unit exchange in surfactin synthetase[45]; deleting or rearranging the order of entire modules, e.g. module deletions and fusion in tyrocidine synthetase[46,47]; and altering the specificity of the A domain in situ, e.g. the L-Glu activating domain of surfactin synthetase was modified to preferentially select L-Gln[48].

As almost every event in nonribosomal peptide synthesis is governed by protein-protein recognition or substrate specific events, alteration of the protein domains or peptide substrates within a biosynthetic complex can have repercussions elsewhere in the assembly line. NRPS reprogramming experiments have already revealed that the positioning of domain or module fusion sites and alteration of the substrate specificity of one domain within a module (either by point mutation or by complete domain replacement) has implications on the productivity of the synthetase and the activity downstream domains.

In the first experiment of its kind, exchange of A-PCP units in surfactin synthetase yielded the expected products but the productivity of the synthetases was drastically reduced[45]. This decrease in productivity was subsequently attributed to the substrate specificity requirements of the C domains[31]. A better understanding of domain borders and the identification and characterisation of the NRPS domain linkers has already been exploited to generate hybrid NRPSs, using module deletion and fusion, with improved yields[47].

The selective association and communication between the individual synthetases of a biosynthetic complex is critical for the synthesis of the predefined peptide. The identification, mutation and deletion of the short stretch of residues, termed short communication-mediating

(COM) domains, at the termini of the interacting tyrocidine synthetases (TycA, TycB and TycC) determined their decisive role in the correct protein-protein recognition of associated peptide synthetases[49]. Further experiments have identified key residues important for maintaining the correct, or preventing the incorrect, interaction between the tyrocidine synthetases. Mutation of these residues has proven successful in switching the specificity of one synthetase for another, has aided the formation of an artificial hybrid NRPS complex and the combinatorial biosynthesis of various designed peptide products[49].

Compared to domain and module swapping the alteration of the substrate specificity of the A domain by targeted mutation of the active site residues is a relatively small modification. To date the majority of changes in A domain substrate specificity have been fairly trivial, i.e. the native and achieved substrates have had side chains of very similar size, overall polarity and shape. Altering an A domain from one that recognizes a large substrate to one that recognizes a small substrate, or vice versa; from one that is specific for a hydrophobic substrate to a hydrophilic substrate, or vice versa; and engineering an A domain with relaxed substrate selectivity capable of utilising a wide range of substrates are the major challenges in this area. An A domain with broad substrate selectivity would help to achieve one of the major goals of the engineered biosynthesis field - the truly combinatorial synthesis of peptides[30]. The possible degree of A domain substrate specificity switching may however be dictated by the substrate specificity imposed by downstream domains, particularly the C domain.

Although the A domains have been studied extensively, knowledge of the selectivity mechanism is still relatively rudimentary. Understanding the molecular basis of this selectivity is critical for informed reprogramming of these domains. Determining the substrate selectivity mechanism for the other NRPS domains is similarly important, as if they are controlled by a relatively minor number of residues the potential to alter them in concert with the A domains may arise. The following aspects of NRP synthesis are as yet unknown and the answers are likely to further the progress of synthetase reprogramming. How do the PCP domains maintain the correct order of interactions when there is a choice of domain partners? Do the linkers have a role in either maintaining the structure of the modules or in

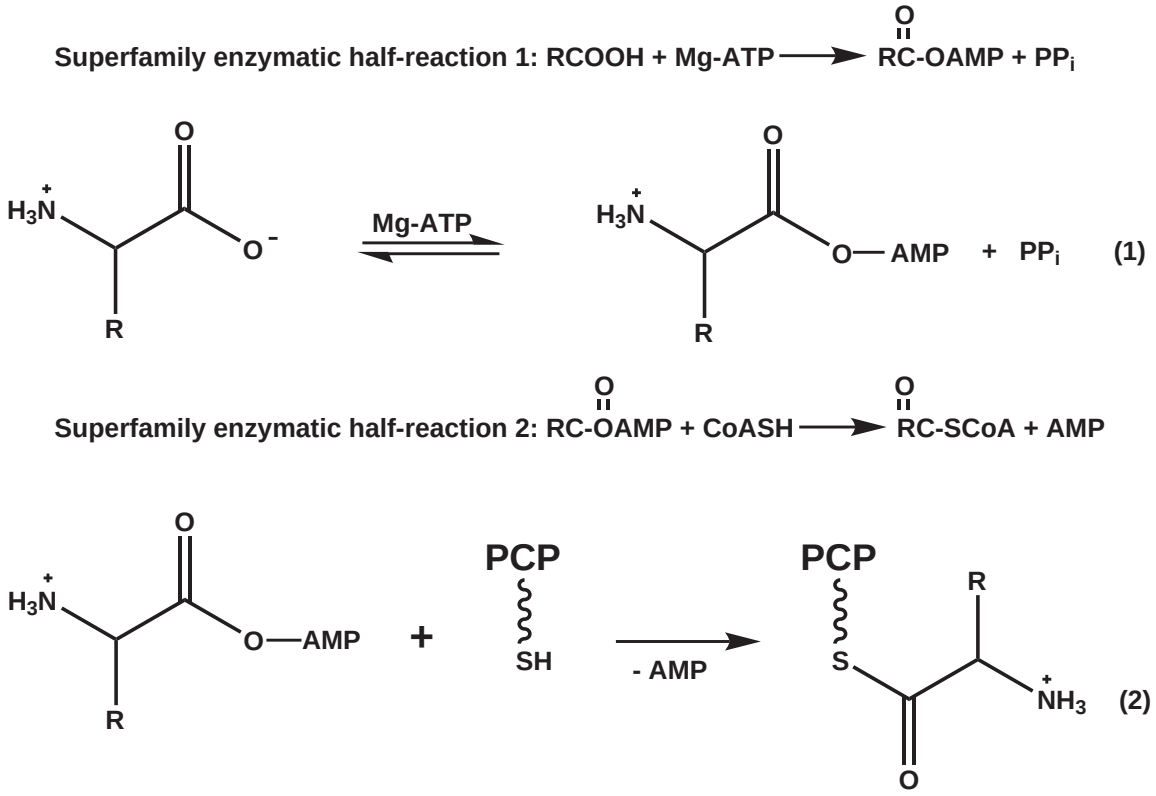


Figure 1.6: **The two half-reactions of the adenylate forming superfamily of enzymes.** Adapted from figure 1 of[54].

aiding domain interactions?[4]

1.4 The Adenylation Domain

The adenylation domains (A domains) of NRPSs; acyl-, acetyl- and aryl-Coenzyme A (A) synthetases/ligases; and insect luciferases are the three subfamilies that constitute the adenylate-forming superfamily of enzymes (PFAM00501) which catalyse two sequential half-reactions via a ping-pong mechanism[9,50–53]. The first half-reaction is the conversion of a carboxylic acid substrate to an acyl-adenylate by consumption of Mg-ATP. In the second half-reaction the acyl-adenylate intermediate is esterified with either CoA (acetyl-, acyl- and aryl-CoA synthetases/ligases) or the enzyme bound CoA derivative PPant (A domains), or oxidised with molecular oxygen (insect luciferases).

The adenylate-forming superfamily of enzymes share between 20 and 40 % sequence homology and contain several conserved motifs. The ten A domain specific conserved mo-

tifs, denoted A1-A10, are shown in table 1.1[7,9,22,50,55]. The superfamily is characterised by a glycine/serine/threonine rich motif (motif A3 in the A domains)[56] that is homologous to the Walker type A motif[57]. The Walker motif forms a traditional phosphate binding loop (P loop) found in all guanosine and some adenosine nucleotide-binding proteins[58–60]. Adenylate-forming enzymes are, on average, 500–700 residues in length and adopt a common fold consisting of a large 400–550 residue N-terminal subdomain and a smaller 100–130 residue C-terminal subdomain[52,61–63].

Enzymes from this superfamily are thought to exploit a "domain alternation" strategy to catalyse the two half-reactions. Comparison of structures co-crystallised with the first and second half-reaction structures revealed a large difference in the orientation of the C-terminal domain relative to the N-terminal domain. The change in C-terminal domain orientation between the two states presents different sets of residues from the smaller domain to the active site. The alternation between the two conformations reconfigures the enzymes single active site enabling catalysis of the two half-reactions[52,53].

1.4.1 A Domain Reaction mechanism

A domains select the amino acid substrate and form a highly reactive aminoacyl adenylate intermediate by reaction with Mg-ATP (half-reaction 1). Following PP_i release, the thiol at the end of the PPant arm attacks the activated aminoacyl group displacing AMP and resulting in the covalent tethering of the substrate to the PPant arm as a thioester (half-reaction 2). The first half of this reaction can be studied by the carboxyl substrate-dependent reversal of adenylation with labelled PP_i[64]. The two stage reaction and mechanism of the A domains can be seen in figure 1.6 and figure 7.1 from appendix 7.1.2 respectively.

In ribosomal synthesis, aminoacyl-tRNA synthetases perform a role equivalent to that of the NRPS A domain, however these enzymes share neither sequence nor structural homology with the A domains[65,66]. NRPSs exhibit moderate substrate specificity compared to the aminoacyl-tRNA synthetases[67–69]. As the A domains specifically select and activate the monomer to be incorporated they are considered the primary determinants of substrate

Core motif [a]	Consensus sequence
A1	L(TS)YxEL
A2 (core 1)	LKAGxAYL(VL)P(LI)D
A3 (core 2)	LAYxxYSTG(ST)TGxPKG
A4*	FDxS
$A5_{aa}$	NxYGPTETTxx
$A5_{aryl}$[†]	QVxFMAEGLVN
A6 (core 3)	GELxJGx(VL)ARGYL
A7 (core 4)	Y(RK)TGDL
A8 (core 5)	GRxDxQVKIRGxRIELGEIE
A9	LPxYM(IV)P
A10	NGK(VL)DR

Table 1.1: **Conserved motifs of the NRPS Adenylation domains[6].** [a] Former nomenclature is given in brackets. * This motif differs in aryl activating domains. [†] A5 motif from aryl activating domains[73]

specificity. Some A domains demonstrate higher substrate specificity than others[24] and for some A domains the substrate incorporated has been shown to depend on those available in the growth media[67–72].

As NRPSs are large multimodular proteins comprised of structurally and functionally independent domains acting in an assembly line manner, A domains usually represent one of the constituent parts of these multimodular proteins. Precise dissection at the boundary of the structural A domain[74,75] has shown excised domains are soluble and catalytically active when expressed heterologously as separate units[24,74,76]. A domains that naturally occur and function as distinct enzymes primarily incorporate aromatic carboxy acids. NRPSs that produce bacterial siderophores, often incorporate aryl acid derivatives at the N-terminal end of the peptide chain. In these synthetases the A domain from the first module that selects and activates the aryl-acid substrate exists as a distinct stand-alone A domain, e.g. DhbE and EntE in bacillibactin and enterobactin synthesis respectively[77].

1.4.2 A domain substrate specificity

The gramicidin S biosynthesis operon from *Bacillus brevis*, shown in figure 1.7, contains the structural genes *grsT*, *grsA* and *grsB*. *GrsA* and *grsB* code for synthetases GrsA and GrsB, respectively. In 1997 the structure of PheA (pdb 1AMU), the phenylalanine activat-

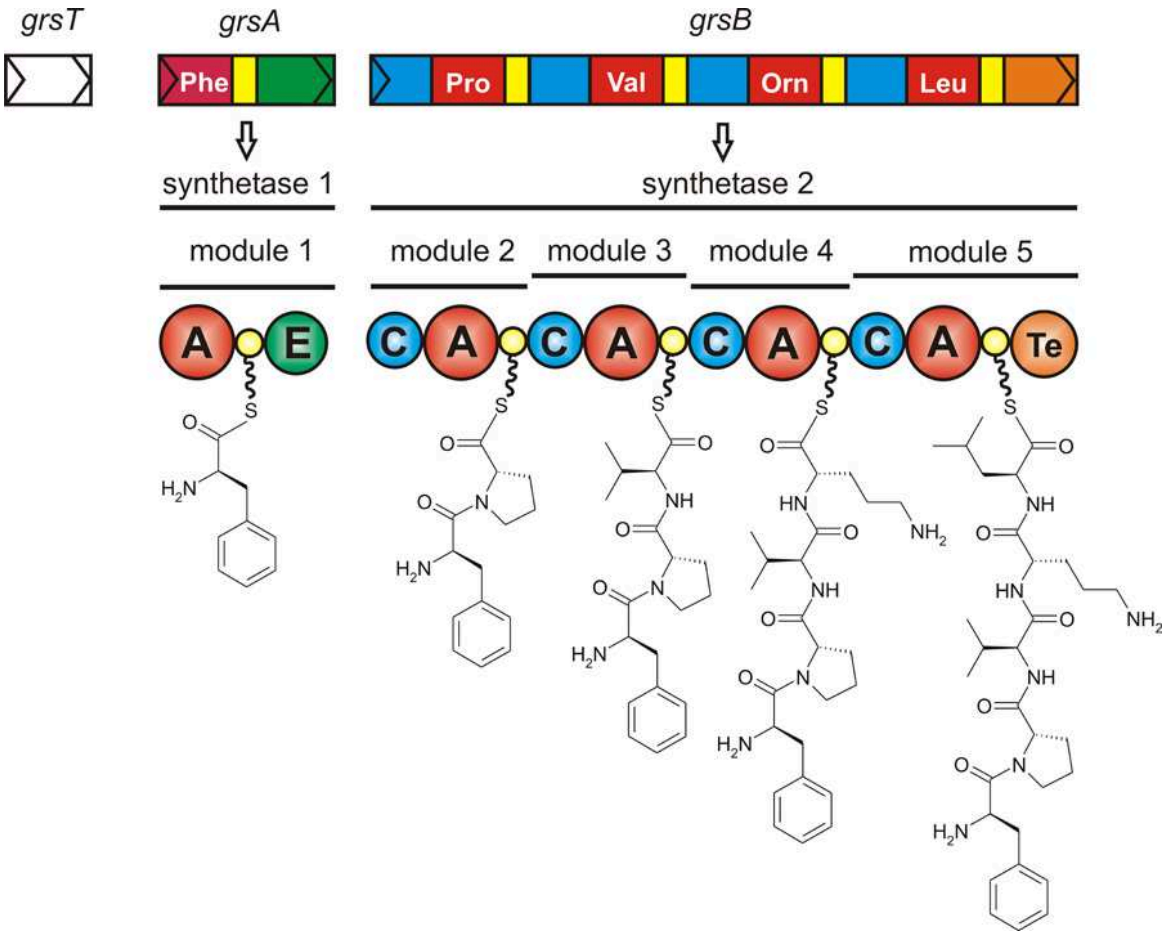


Figure 1.7: **The Gramicidin S biosynthetic gene cluster (*grs*).** GrsA is the initiation module, containing an epimerisation domain that catalyses the inversion of L-Phe to D-Phe. GrsB is composed of four modules that incorporate proline, valine, ornithine and leucine sequentially. Image modified from figure 1 in [78].

ing A domain of GrsA, was determined co-crystallised with L-Phe and AMP; the hydrolysis products of the adenylate intermediate[62]. The 514 residue polypeptide chain folds into two compact domains; a large 412 residue N-terminal domain (A_{core} domain) and smaller 102 residue C-terminal domain (A_{sub} domain). Few direct protein-protein contacts are formed between the domains. The A_{core} domain contains three subdomains: subdomains A and B are both β-sheets, and subdomain C is a distorted β-barrel. These subdomains pack together to form a five-layered $\alpha\beta\alpha\beta\alpha$ tertiary structure. No interpretable electron density was obtained for the highly conserved A3 motif residues. The A_{sub} domain comprises two subdomains, D and E; a small two strand β-sheet (subdomain D) and two helices which pack against one side of a three-stranded anti-parallel β-sheet (subdomain E). The ligands are bound in a cleft at the domain interface that is lined mainly by polar and charged residues.

Determination of the structure of PheA greatly facilitated the study of A domain specificity. Co-crystallisation with L-Phe identified the location of the substrate binding pocket and allowed determination of the residues that line the pocket and make contact with L-Phe. The structure of PheA, and the L-Phe and AMP binding pockets can be seen in figure 3.16. Of the ten PheA L-Phe substrate binding pocket residues, nine (Asp 235, Ala 236, Trp 239, Thr 278, Ile 299, Ala 301, Ala 322, Ile 330, Cys 331) are contributed by the A_{core} domain and are located between, and inclusive of, motifs A4 and A5. D235 and I330 line the top; W239, T278 and I299 the bottom; and A236, A301, A322 and C331 the sides of the PheA binding pocket.

The first residue, D235, is well positioned to form hydrogen bonds with the substrate α-amino group. The tenth residue is the strictly invariant lysine residue (Lys 517) from the A10 motif (K^{A10}) that is contributed by the C-terminal domain and resides on a long loop that projects into the active site. In the PheA structure the K^{A10} residue is well placed to form key polar interactions with both ligands; the α carboxy group of the Phe substrate and the ribose O4′ and O5′ atoms of AMP. The participation of the K^{A10} residue in Mg-ATP binding was biochemically determined by fluorescein 5′-isothiocyanate affinity labelling of TycA of tyrocidine synthetase[79]. The importance of this residue in the first half-reaction was highlighted by the mutation of K^{A10} to Gln in the *B. subtilis* surfactin synthetase (Sr-

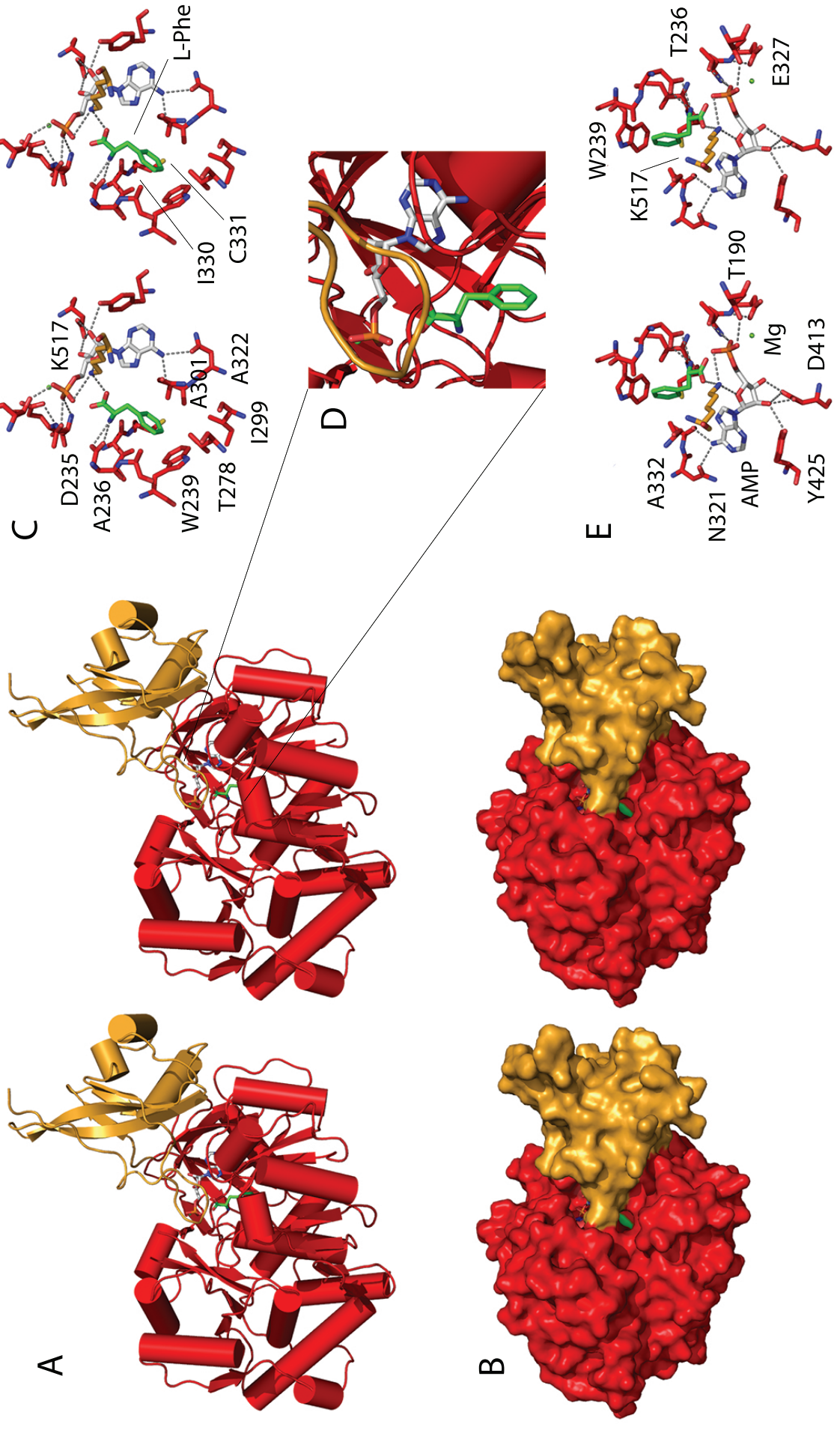


Figure 1.8: **The structure of the PheA.** Shown from two different perspectives using a cartoon (A) and solvent accessible surface representation (B). The N-terminal domain (A_{core}) is in red and the C-terminal domain (A_{sub}) in orange, the phenylalanine ligand in green, AMP in grey and Mg ion in lime. The active site (D) is at the interface between the two domains. The L-phenylalanine active site is shown in C and the AMP binding site in E.

fAB) valine-activating domain which resulted in a 94 % reduction in activity as compared to the wild-type enzyme[80].

Using these structural data and exploiting the relatively high sequence identity of the A domains, two independent bioinformatics studies identified an empirical correlation between the ten residues corresponding to those lining each A domain binding pocket and the substrate activated[81,82]. The profile of ten residues, determined using sequence analysis, for each substrate activating A domain is commonly referred to as the "specificity conferring code". Subsequent analyses of additional A domain sequences have shown that identical substrates can be activated by domains with different predicted selectivity pocket residues[83,84]. This apparent degeneracy in the ten residue specificity conferring code is thought to arise from including residues lining the bottom of the binding pocket in the specificity profiles of the domains that activate the smaller substrates (e.g. proline and threonine) which are thought to utilise only residues lining the top of the binding pocket[30,85]. Determination of the structure of the *Bacillus subtilis* stand-alone A domain DhbE which activates the aryl acid 2,3-dihydroxybenzoate (DHB) enabled refinement of the specificity conferring code for these non amino acid activating domains. The authors' comprehensive study of the three determined DhbE structures: apo (pdb 1MDF), complexed with AMP and DHB (pdb 1MD9), and complexed with the adenylate (pdb 1MDB), in tandem with sequence alignment and modelling studies identified a structural basis for discerning between A domains that activate DHB and those that activate salicylic acid (SAL)[73].

1.4.3 Domain Alternation

Members of the adenylate-forming family have been structurally characterised in three states: without ligands (apo); with substrates, products or analogues of the first half-reaction; or with substrates, products or analogues of the second half-reaction. A representative list of these structures is shown in table 1.2. Comparison of the structures of family members determined in the presence of first and second half-reaction ligands has identified two distinct conformations of these enzymes, which differ in the orientation of the C-terminal domain relative to the N-terminal domain. The determined structures can therefore, be divided

into two groups depending on whether they are in the conformation thought productive for catalysing the first half-reaction (conformation 1) or second half-reaction (conformation 2).

Figure 1.9 shows structures in both conformations. The N-terminal domains have been superimposed, highlighting the difference in C-terminal domain positioning. Apo state enzymes have been determined in both conformational states. While the position of the C-terminal domain relative to the N-terminal domain is equivalent within the structures of the two groups, the degree of rotation of the C-terminal domain varies in the first half-reaction structures and this group can be sub divided into three further groups (called 1.1, 1.2 and 1.3 in table 1.2). The varying C-terminal domain rotation exhibited by the first half-reaction structures does not however affect the residues presented to the active site. Only one residue from the C-terminal domain - the invariant Lys residue from motif A10 located on a long loop that projects into the active site - is critical for binding the substrates of the first half-reaction. This is in direct comparison to the numerous C-terminal domain residues which interact in the second half-reaction with either CoA or the PPant portion of CoA, which are located on the opposite face of the C-terminal domain, in the β-hairpin region of motif A8. The variation observed in the conformation 1 structures may therefore be a direct result of the fewer C-terminal domain residues participating in the reaction and therefore stabilising the conformation.

Conformation 1

The PheA[62] and DhbE[73] A domain structures, *Saccharomyces cerevisiae* acetyl-CoA synthetase (yAcS) structure co-crystallised with AMP (pdb 1RY2)[87], and the *Alcaligenes sp. AL3007* 4-chlorobenzoate:CoA ligase (CBL) apo (pdb 3CW8) and co-crystallised with 4-chlorobenzoate (pdb 3CW9) structures[63], are representative of the conformation thought productive for formation of the adenylate (conformation 1). In this conformation the invariant A10 motif Lys (K^{A10}) residue interacts with both the substrate and AMP molecule. Fluorescein 5′-isothiocyanate affinity labelling of TycA identified K^{A10} as forming part of the Mg-ATP binding site[79]. Mutation of K^{A10} to Gln in the *B. subtilis* SrfAB Val-activating domain resulted in a 94% reduction in activity when compared to the wild-type enzyme[80].

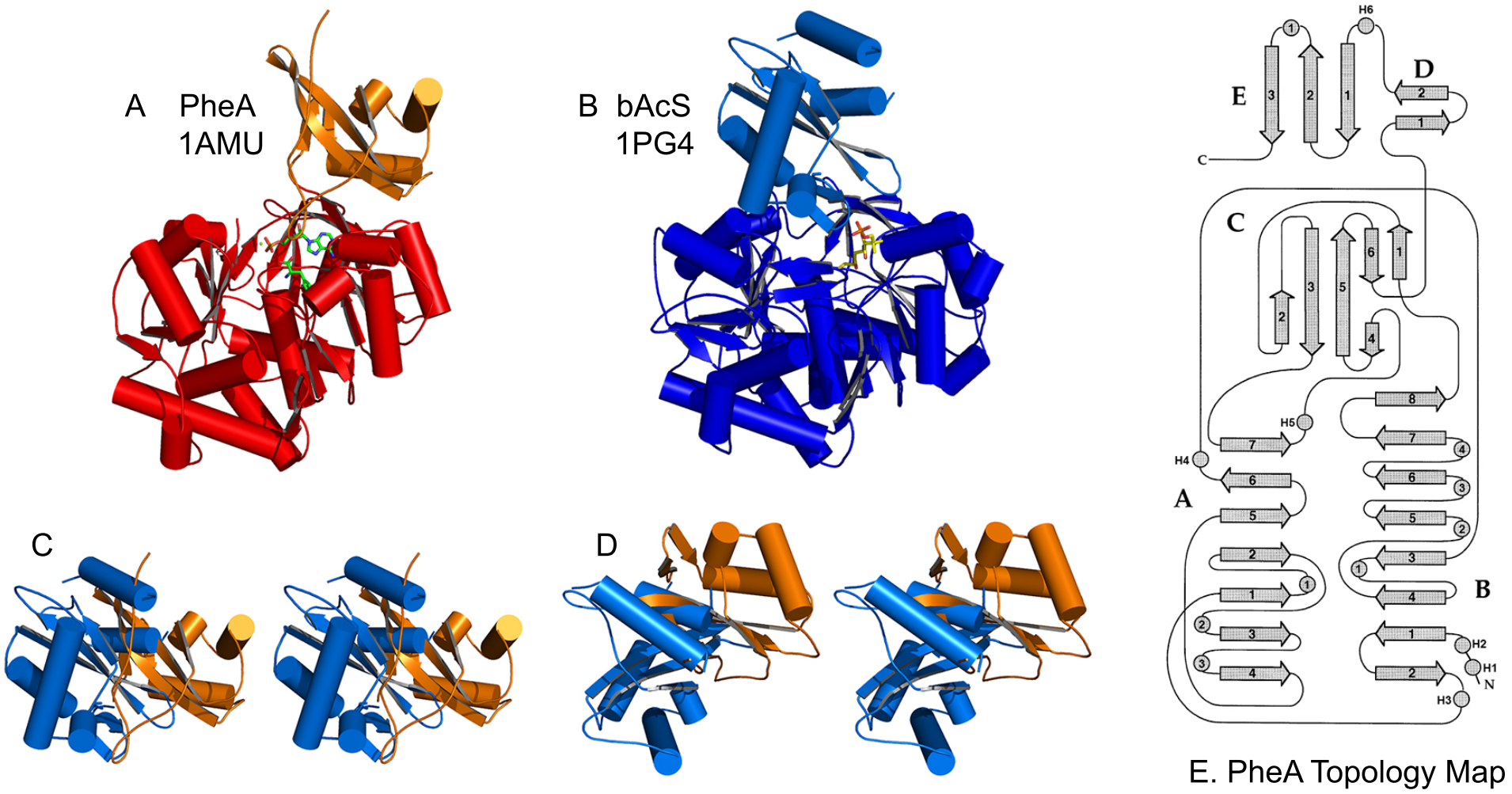


Figure 1.9: **The first and second half-reaction conformers of the adenylate forming superfamily** exemplified by the structures of (A) PheA[62] A domain (pdb 1AMU) in red/orange and (B) bAcS[52] (pdb 1PG4) in blue/teal. The PheA structure is representative of the first half-reaction conformation and the bAcS structure of the second half-reaction conformation. The C-α atoms of the A_{core} domains have been aligned and the orientation of the bAcA A_{sub} domains relative to the PheA A_{sub} domain displayed ijn two orientations, C and D. Figure E shows the topological arrangement of the secondary structural elements in PheA[62].

Conf	Enzyme	Organism	Name	Ligand	Mutation	PDB code
1.1	Luc	*Photinus pyralis*				1LCI[61]
	Luc	*Photinus pyralis*		bromoform		1BA3[86]
1.2	A domain	*Bacillus brevis*	PheA	Phe, AMP, Mg^{2+}		1AMU[62]
	A domain	*Bacillus subtilis*	DhbE			1MDF[73]
	A domain	*Bacillus subtilis*	DhbE	Dhb-Adenylate		1MDB[73]
	A domain	*Bacillus subtilis*	DhbE	Dhb, AMP		1MD9[73]
	AcS	*Saccharomyces cerevisiae*	yAcS	AMP		1RY2[87]
	CBL	*Alcaligenes sp. AL3007*				1T5H[63]
	CBL	*Alcaligenes sp. AL3007*		4-chlorobenzoate		1T5D[63]
	Luc	*Luciola cruciata*		AMP		2D1Q[88]
	Luc	*Luciola cruciata*		Oxyluciferin, AMP		2D1R[88]
	Luc	*Luciola cruciata*		CSO, SLU		2D1S[88]
	Luc	*Luciola cruciata*		CSO, SLU	S286N	2D1T[88]
	LC-FACS	*Thermus thermophilus HB8*	ttLC-FACS			1ULT[89]
1.3	A domain	*Bacillus subtilis*	SrfAC-A	Thr		2VSQ[38]
	MC-ACS	*Homo sapien*		Mg-ATP, unk		3C5E [90]
	CBL	*Alcaligenes sp. AL3007*		4CBA-Adenylate		3CW8[53]
	CBL	*Alcaligenes sp. AL3007*		4CB	D402P	3DLP[53]
2	AcS	*Salmonella enterica*	bAcS	CoA, PRX, EDO, Mg		1PG4[52]
	AcS	*Salmonella enterica*	bAcS	CoA, PRX, EDO, Mg	K609 acetylated	1PG3[52]
	AcS	*Salmonella typhimurium*	bAcS	ACT, CoA, AMP		2P2F[91]
	AcS	*Salmonella typhimurium*	bAcS	PRX	R584A	2P20[91]
	AcS	*Salmonella typhimurium*	bAcS	CoA, PRX	V386A	2P2B[91]
	AcS	*Salmonella typhimurium*	bAcS	CoA, PRX	K609A	2P2J[91]
	AcS	*Salmonella typhimurium*	bAcS	PRX	R194A	2P2M[91]
	LC-FACS	*Thermus thermophilus HB8*	ttLC-FACS	ANP, Mg		1V25[89]
	LC-FACS	*Thermus thermophilus HB8*	ttLC-FACS	MYR, AMP, Mg		1V26[89]
	MC-ACS	*Homo sapien*				3B7W [90]
	CBL	*Alcaligenes sp. AL3007*		4-CP-CoA		3CW9[53]

Table 1.2: **Structures of the adenylate forming superfamily of enzymes.** Luc: luciferase; A domain: Adenylation domain; AcS: acetyl-coenzyme A, CBL: 4-chlorobenzoyl-CoA synthetase, CSO: S-hydroxycysteine, SLU: 5′-O-[N-(dehydroluciferyl)-sulfamoyl]adenosine, LC-FACS: long chain fatty acyl-CoA synthetase, CoA: coenzyme A, PRX: adenosine-5′-monophosphate-propyl ester (adenosine-5′-propylphosphate), EDO: 1,2-ethanediol (ethylene glycol), MYR: myristic acid, ANP: phosphophosphonic acid-adenylate ester, ACT: acetate, MC-ACS: Medium chain acyl-Coenzyme A synthetase, unk: unknown acyl ligand, 4-CP-CoA: 4-chlorophenacyl-CoA.

Examination of the PheA structure reveals the Lys residue is capable of forming hydrogen bonds with both the α carboxy group of the Phe substrate and the ribose O4′ and O5′ atoms of AMP[62].

In the structure of *Salmonella enterica* acetyl-CoA synthetase (bAcS) determined co-crystallised with adenosine-5′-propylphosphate and Coenzyme A (pdb 1PG4), which is representative of the second half-reaction conformation, the equivalent K^{A10} residue is located $\sim 27\mathring{A}$ from the active site[52]. The exclusive importance of this residue in the first half-reaction has been biochemically determined in numerous enzymes of this superfamily. Acetylation of the K^{A10} residue (K609) in *S. enterica* bAcS has been shown to inhibit catalysis of the first half-reaction without affecting the ability of the enzyme to catalyse the second half-reaction[92,93]. Determination of the acetylated form of bAcS showed no conformational changes were induced by acteylation of this residue (pdb 1PG3)[52]. Mutation of this Lys residue in *Photinus pyralis* firefly luciferase (Luc)[94], in propionyl-CoA synthetase (PrpE) from *S. enterica*[95], and *Salmonella typhimurium* bAcS[91] dramatically reduced the ability of the enzyme to catalyse the first half-reaction while having little effect on the ability of the enzyme to catalyse the second half-reaction.

Mapping the A domain motifs onto the PheA structure revealed the majority are located adjacent to the enzyme active site, see figure 1.10. Of the ten conserved motifs biochemical characterisation has indicated a role for residues from eight of the motifs (A3-A10) in the first half-reaction. Motifs A1 and A2 are both far from the enzyme active site and it has been proposed that they are conserved for structural reasons. The A1 motif residues form part of a large helix that links strand β-*B2* to β-*A1*. This helix strongly contributes to the fold of the N-terminal domain[22]. Motif A2 immediately follows strand β-*A1* forming a short helix and strand β-*A2*. The strand β-*A2* residues interact with the A3 motif residues that form strand β-*A5*.

Motifs A4, A5 and A10 all contribute residues to the substrate binding pocket. The second and third residues of motif A4 (F**D**xS); second, tenth and eleventh residues of A5 (N**x**YGPTETT**x**x); and third residue of A10 (NG**K**(VL)DR) together constitute six of the ten binding pocket residues. In PheA the side chain of Asp 235 forms hydrogen bonds to

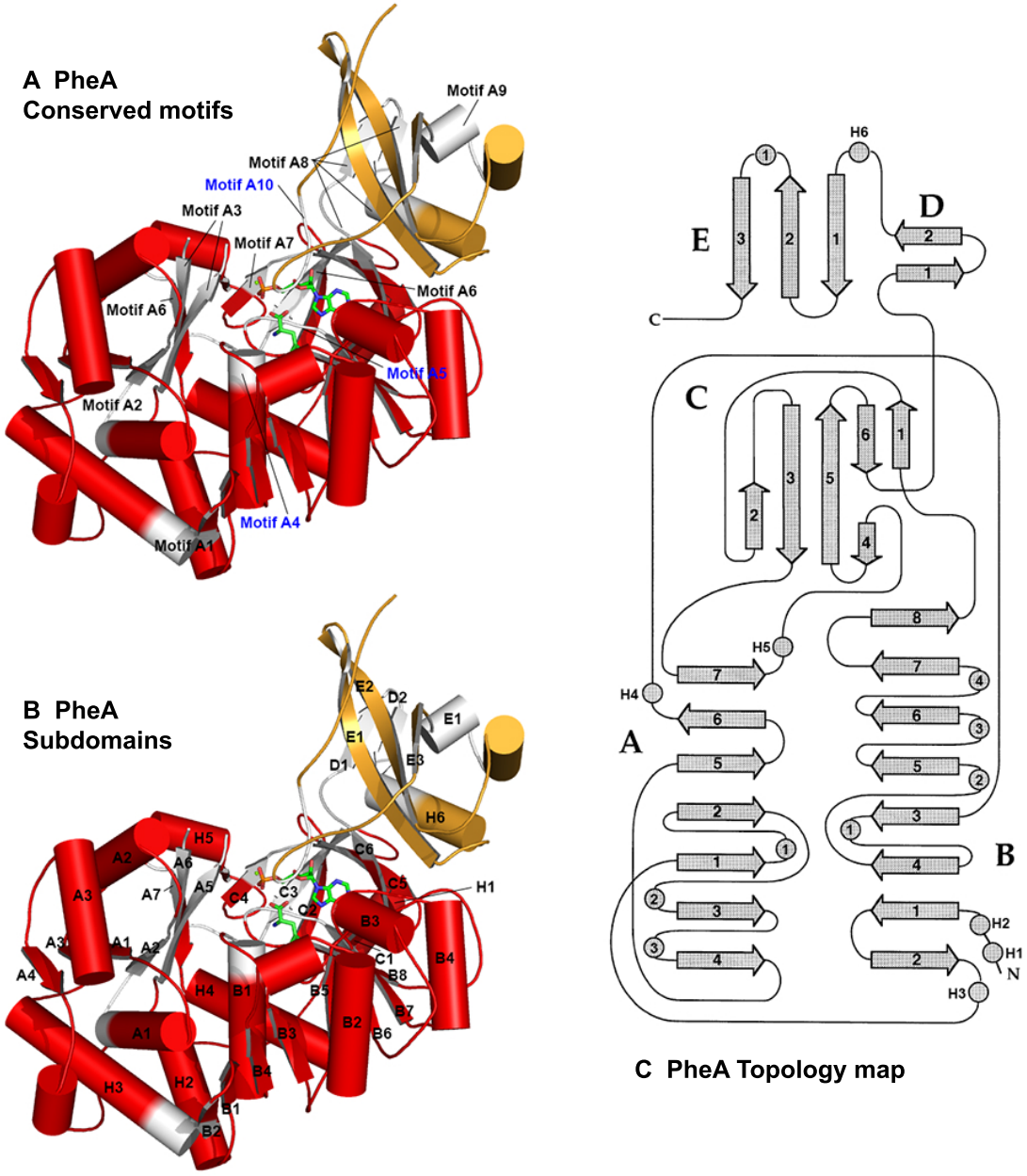


Figure 1.10: **PheA conserved motifs, subdomains and topology map**. A) The conserved A domain motifs mapped onto the structure PheA[62]. The A_{core} domain is in red, the A_{sub} domain in orange, phenylalanine, AMP and Mg ligands in green B) PheA structure annotated with subdomains and secondary structural elements. C) Topological arrangement of the secondary structural elements in PheA.

the α amino group of the Phe substrate[62]. This aspartic acid residue is only invariant in amino acid activating A domains, in DhbE the neutral amino acid Asn replaces this Asp residue[73].

In the first half-reaction conformation of *A. sp. AL3007* CBL, His 207 forms a hydrogen bond to the oxygen atom bridging the 4-CB and AMP portion of the adenylate. His 207 is the first residue of the A4 motif, the equivalent residue in PheA is Phe 234. The position of H207 in conformation 1 structure occludes the PPant arm thiol binding site thus preventing binding of CoA to the wrong conformation. In the conformation 2 structure Glu 410 rotates into the active site interacting with His 207 and pulling it from the active site. The thiol of the PPant arm binds in the space vacated by His 207[53].

In PheA, as well as lining the Phe substrate binding pocket, additional A5 motif residues are well placed to interact with the AMP ligand. The main chain carbonyl of A322 may accept a hydrogen bond from the amino group of the adenine base and the main chain carbonyl oxygen of Gly 324 may accept a hydrogen bond from the α-amino of the Phe substrate[62]. The product of a *B. brevis* Nagano E-4 strain mutant gene was found to contain a mutation of the A5 motif glycine residue (G1793D) in the valine activating A domain of gramicidin synthetase 2 (GrsB)[96] which was responsible for abolishing enzyme activity[97].

Residues from motifs A3, A6, A7, A8 and A9 have been biochemically determined to be important for binding the Mg-ATP substrate or forming the amino acyl-adenylate. As the core 2 (A3) motif defines the superfamily, the function of numerous motif residues have been investigated. The A3 motif is located in a disordered loop region connecting the anti-parallel strands β-*A5* and β-*A6*. These strands are flanked by strands β-*A2* and β-*A7* formed by the residues of motifs A2 and A6 respectively. In the P loop[57,98,99] the conserved Lys residue aids binding of the ATP γ-phosphate atoms, by analogy a similar role was suggested for the invariant A3 lysine residue (LAYxxYSTG(ST)TGxP**K**G), K^{A3}. Mutation of this conserved Lys to Arg and Thr in TycA reduced the enzyme activity to 90% and 99.5% that of the wild type enzyme respectively[100]. In an independent study the mutation of K^{A3} to Arg in TycA resulted in a 75% reduction in activity when compared to the wild-type enzyme[75]. Mutation of the first Lys (LAYxxYSTG(ST)TG**x**PKG) to Gln in the *B. subtilis*

valine activating domain of surfactin synthetase 2 (SrfAB) had no significant effect on enzyme activity (activity was reduced to 91% of the wild type); mutation of the conserved Lys to Gln however, reduced activity of the enzyme to 39% that of the wild-type enzyme[80]. In 4CL from *Arabidopsis thialana* mutation of K^{A3} to Ser reduced enzyme activity to 3% that of the wild type enzyme[101]. Separate mutation of each of the three A3 motif Gly residues to Ala (YST**G**(ST)T**G**xPK**G**), and Pro to Val (LAYxxYSTG(ST)TGx**P**KG) in the *B. brevis* TycA[102] Phe activating A domain[103] had no significant effect on the adenylation activity of the enzyme[100]. Direct participation of the first Gly residue (YST**G**(ST)TGxPKG) in the adenylation reaction was demonstrated in the Val activating A domain of GrsB, however, when a mutant with a point mutation of this residue to Asp was found to be completely inactive[97]. Mutation of the core 2 loop residues G163, G166, P168 and K169 (LAYxxYST**G**(ST)T**G**x**P**KG) in *Pseudomonas sp. CBS3* CBL resulted in impaired catalysis of the CBA adenylation partial reaction[104]. In the PheA complex the side chain of the first Thr residue (LAYxxYS**T**G(ST)TGxPKG) is well placed to form a hydrogen bond to the α-phosphate oxygen atom[62]. Although residues ^{192}GTTGN^{196} of this motif - which would form the loop - were not determined in the PheA structure, the orientation and proximity of these residues to the AMP binding site suggest an interaction with the PP_i leaving group[22].

The structure of human medium chain acyl-CoA (MC-ACS) was determined in complex with Mg-ATP (pdb 3C5E) by the Structural Genomics Consortium in Toronto[90]. This MC-ACS structure represents the first superfamily enzyme co-crystallised with ATP and provides insight into the role of the A3 motif and additional residues in Mg-ATP binding. The A3 motif residues that form a hydrogen bonding network with the α β and γ-phosphate oxygen atoms are shown in italic: ^{215}LAYxxY**T**SG(T)SGxP**K**G^{230} - where the TSG(ST)S sequence replaces STG(ST)T of the standard core 2 motif. The hydroxyl side chain of S222 interacts with the α-phosphate atoms. The amino backbone and hydroxyl side chain groups of T224 interact with the β-phosphate oxygen atoms. The hydroxyl side chain groups of T221 and S225, amino backbone groups of G223 and S225 and amino side chain group of K229 interact with the γ phosphate oxygen atoms. Additionally an oxygen from each of the β- and γ-phosphates of ATP is coordinated to the Mg^{2+} which is also coordinated to

four water molecules, two of which interact with the A5 Glu residue (NxYGPT**E**TTxx).

Motif A6 forms strands β-*A7* and β-*C3* of the N-terminal subdomain. Photoaffinity labelling of the *B. brevis* tyrocidine synthetase 1 (TycA)[102] Phe activating A domain[103] with 2-azidoadenosine triphosphate (2-azido-ATP) identified residues (^{373}**G**YWWRPDLTAEK^{384}) which include a region of this motif, thus indicating its involvement in catalysing aminoacyl adenylate formation[105].

In PheA the A7 motif residues link strands β-*C4* and β-*C5* and are adjacent to both the adenine binding site and the conserved Arg residue of motif A8. This motif bears homology to the ATPase motif[106–110] which plays a role in nucleotide binding[108]. Mutation of the invariant Asp residue (Y(RK)TG**D**L) from A7 motif to Asn and Ser in TycA decreased the phenylalanine-dependent ATP-PP_i exchange activity to 78% and 12% that of the wild-type level respectively[100]. In PheA the position of the Asp side chain enables acceptance of hydrogen bonds from the 2′ and 3′ ribose hydroxyl groups[62].

In the PheA structure residues from the the A8 motif link the A_{core} and A_{sub} domains and form the β-hairpin[111] that is subdomain D of the A_{sub} domain. Sequencing of a mutant *B. brevis* Nagano BII-3 strain gene coding for the Pro activating A domain of GrsB defective in Pro activation identified a point mutation of the second A8 motif Gly (GRxDxQVKIR**G**xRIELGEIE) to Glu. Further mutation of this residue to Ala, Val, Arg and Trp resulted in scarcely active enzymes. These results suggest this residue is essential for aminoacyl-adenylation[112]. Additionally, photoaffinity labelling of TycA with 2-azido-ATP and fluorescein 5′-isothiocyanate indicated this region was involved in catalysing aminoacyl adenylate formation[105] and that the Lys residue (GRxDxQV**K**IRGxRIELGEIE) is involved in Mg-ATP binding[79].

Mutation of the first Arg in motif A9 (LP**x**YM(IV)P) to Thr in TycA resulted in profound loss of activity[113]. The A9 motif residues connect strands β-*E2* and β-*E3* and adopt a helix conformation. These residues are located on the opposite face of the A_{sub} domain to the active site and motif A10. Photoaffinity labelling of TycA with 2-azido-ATP identified the following sequence ^{483}LPAYM**L**PSYFVK^{494} which contains motif A9 indicating the

involvement of this motif in aminoacyl adenylate formation catalysis[105].

Conformation 2 Structures

The *Salmonella enterica* bAcS structures cocrystallised with adenosine-5′-propylphosphate and Coenzyme A[52], *Salmonella typhimurium* bAcS structures determined with various ligands[91], and *A. sp. AL3007* CBL structure co-crystallised with 4-chlorophenacyl-Coenzyme A (4-CP-CoA)[53], are representative of the conformation of the enzyme used to catalyse the thioester forming second half-reaction (conformation 2). Re-orientation of the C-terminal domain in conformation 2 removes the K^{A10} residue from the active site positioning it $\sim$27Å away from the domain interface binding pocket. The alternate C-terminal domain orientation relative to the N-terminal domain presents the A8 motif residues to the active site[52]. Analysis of the second half-reaction structures coupled with the results of biochemical experiments has identified A5 and A8 motif residues as important for the thioester-forming half-reaction.

In the *S. enterica* bAcS second half-reaction structure (pdb 1PG4) the binding site for adenosine-5′-propylphosphate, a mimic for the acyl-adenylate intermediate, is almost completely buried. The position of the AMP moiety and propyl group of adenosine-5′-propylphosphate is comparable to that of AMP and Phe, respectively, in the PheA structure. The nucleotide portion of CoA binds on the surface of the protein and the pantetheine moiety, which is less well-ordered than the nucleotide moiety, passes through a channel between the two domains and points into the AMP binding site. In this structure the A5 motif Glu (E417) residue (NxYGPT**E**TTxx) forms a salt bridge with the third A8 motif Arg (R526) residue (GRxDxQVKIRGx**R**IELGEIE) stabilising the conformation 2 structure[52].

This A5 motif Glu residue also has a role in the first half-reaction as illustrated in the recently determined MC-ACS structure complexed with Mg-ATP and an unknown acyl ligand (pdb 3C5E). This structure provides insight into the positioning of Mg^{2+} and the role of the A5 motif Glu residue in the coordination of this ion. An oxygen from each of the ATP β- and γ-phosphates is coordinated to the Mg^{2+} which is also coordinated to four

water molecules, two of which interact with the A5 motif Glu residue[90]. Mutation of the A5 motif Glu to Gln in CBL from *P. sp. CB53* resulted in a 50-fold reduction in both overall enzymatic activity and reactivity of the first half-reaction[104].

Analysis of the second half-reaction structures has identified a number of residues that interact with the CoA substrate in the thioester forming conformation. Of direct relevance to the A domains are the residues of the A8 motif determined to interact with the PPant arm of CoA. In bAcS the main chain carbonyl atoms of residues S523 and G524, equivalent to the second Arg and Gly residues (GRxDxQVKI**R**GxRIELGEIE) of the A8 motif, form the loop of the β-hairpin and interact with the two amines of the PPant group[52]. In conformation 2 the A8 loop residues occlude the region in which the β - and γ-phosphates of ATP bind in conformation 1. This hindrance of the phosphate binding site prevents ATP binding to the thioester forming conformation[53].

In the *Salmonella typhimurium* bAcS structure (pdb 2P2F) the pantetheine moiety of CoA is less well ordered than the nucleotide portion. In this structure the β-alanine group of CoA passes below the C-α atom of G524 ($G_2{}^{A8}$). This residue was mutated to Ser and Leu - amino acids with increasingly larger side chains. The mutants were subjected to steady-state kinetics to determine kinetic constants for ATP and CoA and, as appropriate, examined using the PP_i-exchange assay. While kinetic constants for ATP were not affected by the G524S mutation, they were for CoA: the k_{cat} for CoA was reduced by a factor of 2, and the k_{cat}/K_M reduced by a factor of 20. Activity of the G524L mutant was undetectable using steady-state kinetics. The G524L mutant revealed wild type enzyme activity levels with the PP_i-exchange assay, yet no detectable activity with the NADH consumption assay, indicating mutation of G524 disrupts the second-half reaction only, presumably by occluding the PPant tunnel[91]. Point mutation of A8 motif residues K445 ($K_1{}^{A8}$) and K457 ($K_2{}^{A8}$) in At4CL2 reduced the overall rate of the partial reaction 2 product caffeoyl-CoA by 96–99%[101]. In Luc from *P. pyralis*the residues equivalent to $R_2{}^{A8}$ (K445 - equivalent to K457 in At4CL2) and $G_2{}^{A8}$ (G446) from the A8 motif were mutated to Gln and Ile respectively. Both mutants exhibited near normal (wild type enzyme) rates of adenylate formation. While Luc does not require CoA for the oxidative reaction that produces light, Luc can utilise CoA

to synthesize L-SCoA (luciferase-SCoA) from L-AMP. The addition of CoA can double the total light output as compared to wild type Luc as production of L-SCoA is accompanied by the release of free luciferase. The K445Q and G446I mutations diminished and abolished, respectively, the enhancing effect of CoA on bioluminescence suggesting that these mutations disrupt binding of the pantetheinyl moiety of CoA[54]. Mutation of R437 - the residue that proceeds the $R_2{}^{A8}$ and $G_2{}^{A8}$ residues - to Asp in the *Escherichia coli* stand alone A domain EntE rendered the enzyme severely compromised for catalysis of the complete reaction yet competent for catalysis of the adenylate[114]. Thus A8 motif residues that form the loop of the β-hairpin have been shown to be involved specifically in the second partial reaction in all three subfamilies of the adenylate-forming superfamily.

The first Asp residue of the A8 motif (GRx**D**xQVKIRGxRIELGEIE) has been determined as the hinge about which domain alternation occurs. The torsional angles of this hinge residue differ in the first and second half-reaction conformations. The ϕ and ψ torsion angles of the Asp residue are: $-60\,^\circ$ and $-32\,^\circ$ for the PheA hinge (D430)[62], and $-74\,^\circ$ and $-29\,^\circ$ for the CBAL hinge (D402)[53], in the first half-reaction structures; $-103\,^\circ$ and $-169\,^\circ$ for the bAcS hinge (D517)[52], and $-90\,^\circ$ and $-164\,^\circ$ for the CBAL hinge (D402) in the second half-reaction structures[53]. The equivalent region is disordered in the *Pontius pyralis* unligated firefly luciferase (Luc) structure[61] indicating inherent flexibility.

Limited proteolysis of TycA digested both the apo and holo-1 state A domain at Arginine 416 from the A8 motif (G**R**xDxQVKIRGxRIELGEIE) indicating flexibility in this region. Digestion of TycA at this residue was reduced in the presence of the adenylation reaction substrates, indicating they protect the enzyme and reduce the flexibility of the enzyme in this region. Additionally mutation of this Arg residue profoundly reduced activity of the enzyme[113].

Mechanism of Alternation

Alternation between these two conformations has been proposed as a strategy to reconfigure a single active site to perform two different reactions. Determination of *A. sp. AL300*

4-chlorobenzoyl-CoA synthetase (CBL) structures in complex with 4CBA-Adenylate and 4-chlorophenacyl-CoA represents the first instance of a single enzyme determined in each of the conformational states[53], though these structures have not been released to the PDB. The original domain alternation theory proposed that binding of CoA (or pantetheine), after formation of the adenylate and dissociation of PP_i, triggered the conformational change from the adenylation-forming conformation to the thioester-forming conformation[52]. While all subsequently determined holo state structures supported this theory, the variation of conformations exhibited by apo state family member structures did not. As such this hypothesis was updated to include the postulation that in the absence of ligands there is an equilibrium between the two conformational states that differs for different proteins within the family, possibly depending on the local environment, and that members may have a preferred crystallisation state. The binding of ligands would induce one or the other of the two conformational states primarily because of steric conflicts in the wrong conformation - in conformation 1 the CoA thiol binding site is obstructed by the second residue of motif A4 (F**D**xS) and the switch from conformation 1 to 2 requires the dissociation of PP_i to free space into which the β- hairpin A8 motif loop residues move[53,115]. Structural and biochemical data support the theory that the two half-reactions proceed via a ping-pong mechanism in which the two independent steps are catalysed by the two observed conformations separated by the domain rotation[53]. This theory also explains the conservation of the A8 and A10 motifs which are on opposing faces of the C-terminal domain approximately 30Å apart.

NRPS A Domain - Evidence for Domain Alternation

Although the A domain has not been determined in the thioester-forming conformation, the similarities between members of the adenylate forming superfamily suggest exploitation of an equivalent domain alternation strategy. In addition to catalysing similar reactions, all superfamily members contain highly conserved motifs and adopt a conserved fold. Pairwise alignment of the following six adenylate forming family members - *Photinus pyralis* Luc, PheA, DhbE, *Salmonella enterica* bAcS, *Saccharomyces cerevisiae* yAcS, and *A. sp. AL3007* CBL - revealed that the superfamily members are not more conserved in sequence

identity or structural homology within a single subfamily than between different subfamilies[63]. The K^{A10} residue that forms key interactions with both the substrate and AMP in the structures of PheA and DhbE has been demonstrated as critical exclusively in the first half-reaction in other subfamilies of this superfamily[91,94,95]. Furthermore the A8 residues critical for binding the pantetheine portion of CoA in the second half-reaction structures of bAcS and CBAL are highly conserved in the A domains. Limited proteolysis studies of tyrocidine synthetase 1 (TycA)[75,113] indicated intrinsic flexibility of the protein at the inter-domain hinge region that is reduced in the presence of the first half-reaction ligands. In PheA the hinge residue (Asp 430) displays torsional angles very similar to those observed in those family members in the adenylate-forming conformation. Together this evidence suggests the A domains exploit a similar strategy of domain alternation to reconfigure the single active site.

The structure of the *B. subtilis* leucine activating A domain from SrfA-C provides insight into the feasibility of A domain alternation within an NRPS module. This A domain, co-crystallised with Leu, incorporates the final monomer for surfactin synthesis in *B. subtilis* and was determined as part of the four domain termination module (C-A-PCP-TE) SrfA-C[38]. In this structure the C domain and A_{core} (N-terminal) domain form a platform upon which respectively the PCP and A_{sub} (C-terminal) domain reside. This platform potentially provides a stable surface for domain reorientation. Although the relative orientation of the A_{sub} domain means the structure is productive for catalysing the adenylate forming half-reaction, the K^{A10} residue is not in the active site. Instead the entire loop is lifted out of the active site and the K^{A10} residue is $\sim$15Å from the substrate. Without the stabilisation of the K^{A10} residue the Leu substrate is bound at the top of the substrate binding pocket in a different orientation to that observed for Phe in PheA. A comparison of orientation of the A_{sub} domains of DhbE, PheA and SrfA-C is shown in figure 1.11.

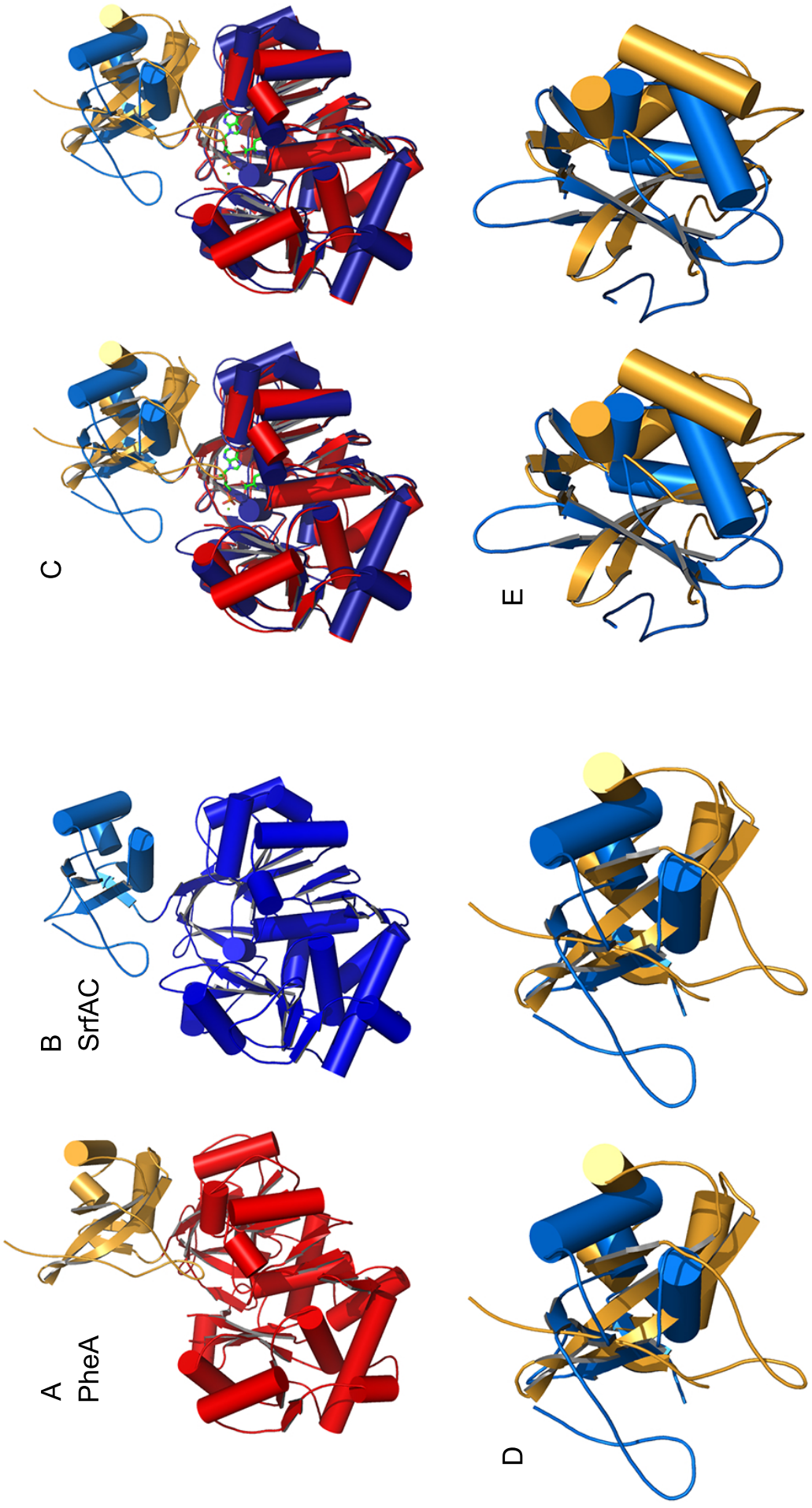


Figure 1.11: **The structures of A domains PheA and SrfA-C**; (A) PheA [62] (pdb 1AMU) in red (A_{core} domain) and orange (A_{sub} domain) and (B) SrfA-C [38] (pdb 2VSQ) in blue (A_{core} domain) and pale blue (A_{sub} domain). (C) The C-α atoms of the A_{core} domains have been aligned and the orientation of the SrfA-C A_{sub} domain relative to the PheA A_{sub} domain displayed in D and E.

1.5 The Peptidyl Carrier Protein Domain

The PCP and aryl carrier protein (ArCP) domains of NRPSs are related in structure and mechanism to the acyl carrier protein (ACP) domains of polyketide synthetases (PKS) and fatty acid synthetases (FAS). These carrier proteins (CP) are small, 80–100 residues, non-catalytic domains that contain a central invariant serine residue[116] located in the shared conserved sequence motif Gx(D/H)**S**(L/I)(D/K)[116,117] and the PCP specific motif LGG(D/H)**S**L[6,116,118]. This invariant serine residue serves as the attachment point for the phosphopantetheinyl moiety (PPant) of coenzyme A[27] which is transferred to the CP by a member of the superfamily of enzymes called phosphopantetheinyl transferases (PPTases)[118]. The mechanism of post-translational modification of CPs by PPTases can be seen in figure 7.1 in appendix 7.1.2.

Two types of PPTase have been identified in bacteria: the AcpS-type which solely activate carrier proteins from primary metabolism (FAS)[119–122] and the Sfp-type with broad substrate tolerance, predominantly and preferentially activating CP of secondary metabolism unless a primary metabolic PPTase is absent[123–125]. The crystal structure of Sfp was solved in 1999 by Reuter and co-workers (pdb 1QR0) see figure 7.2 in appendix 7.1.2. Mutation of the invariant serine to glycine and alanine in the PCP domains of the D-Phe activating module of *B. brevis* ATCC 8185 TycA prevented formation of the thioester bond and tethering of the substrate to the PCP domain[100]. Post-translational modification of PCP, ArCp and ACP converts the domain from its inactive *apo* form to its active *holo* form[118,126,127]. The presence of multiple PPant carrying domains in an NRPS (one PCP domain per module), as opposed to a single central carrier protein, was determined by Stein *et al.* in 1994. Each activated amino acid substrate presented as a thioester on a unique PPant group is central to the proposed multiple-carrier thiotemplated mechanism of non-ribosomal peptide synthesis[27,28].

In NRPSs the second half-reaction carried out by the A domain covalently tethers the aminoacyl adenylate to the terminal thiol of the PPant arm by the formation of a thioester bond[116]. While tethered to the PCP domain the substrate can be modified by optional edit-

ing domains located in the same module (*n*) downstream of the PCP domain. Following optional editing of the substrate the C domain of module$_{n+1}$ catalyses peptide bond formation between substrates covalently bound to the PCPs of the two adjacent NRPS modules (*n* and *n+1*). This reaction results in covalent attachment of the peptidyl product to the PCP domain module$_{n+1}$ and the release of the sulfhydryl group from the PPant moiety of the PCP domain from module *n*.

Covalent attachment of the activated substrates to the PCP domains protects the substrate from bulk solvent, stabilises the reactive intermediates and sequesters the substrate away from competing cellular processes [128]. This latter feature is of particular importance as many of the precursors of nonribosomal peptide synthesis are diverted from primary metabolic processes, or used within other secondary metabolic processes, including by PKSs or other NRPSs [129]. The PCP-PPant-substrate complex commands a higher level of recognition from interacting domains than the growing peptidyl chain would alone [130]. As such this PCP facilitated method of substrate delivery, termed "substrate channelling", increases the overall turnover rate of the NRPS assembly line process. The PCP domain can therefore be thought of as a peptide shuttle specifically communicating with numerous partner domains: upstream A and C domains; downstream C domains; discrete type II TE domains which are responsible for the regeneration of misprimed PCP domains [131] (for the mechanism of type II TE domains see figure 7.7 in appendix 7.1.2); optional modifying domains (including E, Mt and Ox domains) within the module; and a terminal domain that can either be a Te, Red or rarely a C domain.

The NMR solution structure of TycC3-PCP, the third module PCP domain of the *B. brevis* tyrocidine synthetase 3 (TycC3), was determined in 2000 [132]. The PCP domain is a distorted four-helix bundle (αI–αIV) with an overall fold that is common to all NRPSs, FAS and PKS carrier proteins determined to date [37,38,114,119,132–140]. Helices αI and αII of TycC3-PCP are linked by an extended loop and dominate the overall structure as they are longer than helices αIII and αIV. Helix αIII is perpendicular to helices αI and αII and the short turn linking αIII to αIV positions αIV virtually anti-parallel to αI. The conserved serine residue (S45) is located within a stretch of seven flexible residues at the C-terminal of the region linking

helices αI and αII. The same nuclear Overhauser effect (NOE) spectra were obtained for both unmodified and phosphopantetheinylated PCP suggesting the PPant arm is flexible and extends into the bulk solvent, not interacting with the PCP domain[132].

Re-examination of the NMR data acquired during the determination of the TycC3-PCP structure revealed TycC3-PCP in three stable distinct conformational states. These are a unique apo conformation (state A or PCP_A); a distinct holo conformation (state H or PCP_H); and one conformation (state A/H or $PCP_{A/H}$) that is structurally identical to both the apo and holo forms of PCP - with the exception of the attachment of the PPant arm[141]. The state A/H conformation most closely resembles the classic four helix bundle structure identified as the original TycC3-PCP structure. The three states of the TycC3-PCP domain can be seen in figure 1.12.

The unique A state of the apo PCP domain is the most flexible and extended structure. The length of helices αI, αII and αIV are reduced, compared to the A/H state, and a kink is present in helix αII at position Q54. Helix αIII is absent and the residues instead form a long loop (LIII) that is embedded within the protein core between helices αII and αIV. The critical role of the active site serine in the conformational diversity of apo PCP was demonstrated by MD simulations of the A state PCP conformation; the TycC3-PCP_{S45A} mutant exhibited a loss of structural heterogeneity between the two apo-PCP conformations effectively rendering the protein frozen in the A state. Helix αIII is unravelled and extended in the H state conformation of holo PCP. This rearrangement moves helix αIV parallel to helix αI by $\sim$3Å, and relocates helix αII and subsequently the residues linking helices αI and αII and therefore the active site serine residue. A combination of molecular modelling and analysis of the NOE spectra obtained for the holo TycC3-PCP structures (A/H and H conformations) positioned the cofactor as being close to the N-terminus of PCP in the A/H state and the C-terminus in the H state conformation. This indicates that the transition between the A/H and H states of holo TycC3-PCP results in migration of the PPant cofactor across the face of PCP which shifts the terminal thiol group by $\sim$16Å. These data provides direct evidence of the intrinsic flexibility of the PCP domains and that the PPant arm "swings" during non-ribosomal peptide synthesis. The structures of the three states of TycC3-PCP

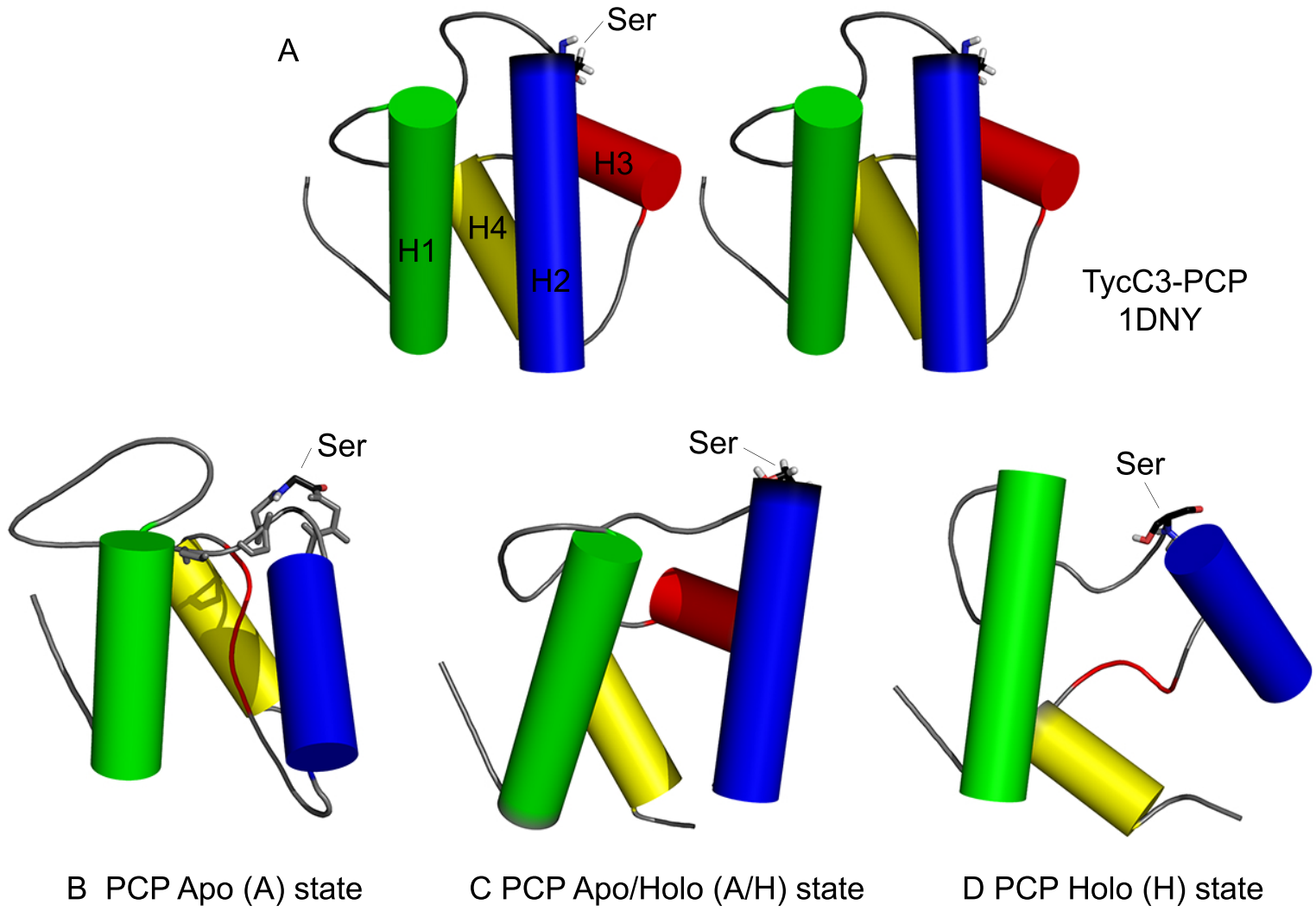


Figure 1.12: **The average NMR solution structures of PCP domain TycC3.** A) the originally determined structure (pdb 1DNY)[132] - the classic four helix bundle CP structure, B) A state apo PCP (pdb 2GDY)[141] - recognised exclusively by PPTase, C) A/H state PCP (pdb 2GDW)[141] - shared by both apo and holo PCP and the structure that most resembles the classic four helix bundle CP structure, and D) H state PCP (pdb 2GDX)[141] - adopted only by holo PCP and recognised by type II thioesterases (TEII). Helix α1 is coloured green, α2 in blue, α3 in red, α4 in yellow, the loops linking the helices (L1, L2 and L3) in grey and the active site serine residue in black.

can be seen from differing angles in figure 7.3 in appendix 7.1.2.

Additional NMR titration experiments identified the residues and specific conformations of TycC3-PCP that interact with the PPTase enzyme Sfp and type II thioesterase SrfTEIII. Sfp was determined to selectively interact with the apo PCP protein, making contacts with the 25 residues that form helix αII and the proceeding αI αII linker - a region accessible in the A state conformation yet not in the A/H state. An interaction surface formed by the C-terminal end of helix αII and the αI–αII and αII–αIII linker regions exclusively in the H state conformation, yet not in the A/H state, was recognised by SrfTEIII.

Interaction points for other partner domains have been identified in various PCP domains using genetic engineering and by crystallising multi-domain PCP containing structures. Replacement of the TycC3 helix αII with the equivalent helix from bacterial FAS ACP enabled the hybrid protein (hTycC3) to be recognised by AcpS, in contrast to wild type TycC3-PCP which is not recognised by AcpS. hTyC3-PCP retained the ability to be recognised by the upstream A domain while recognition by the downstream E domain was abolished. This suggests residues for upstream A domain recognition reside outside helix αII while those that participate in the E domain interaction are present on helix αII[122]. Further investigation of the PPTase and PCP helix αII interaction residues in TycC3-PCP identified that residue K47, which is located two residues downstream of the active site serine, specifically interacts with D40 of Sfp[142]. The PCP residues that specifically interact with the EntD and Sfp PPTases have been determined for the EntB ArCP and EntF PCP. Residues G242 (S -3)[143] and D244 (S -1)[114,143] of the αI–αII loop of EntB-ArCP were determined to interact with EntD and Sfp and residue L1007 (S +1) of helix αII of EntF with Sfp[144]. All of these residues are adjacent to the active site serine residue.

The PCP domain in the TycC PCP_5-C_6 didomain crystal structure is in the A/H conformation suggesting it is primed to interact with the upstream A or C domains. Residue E56 from helix αII was identified at the interface between the domains forming salt bridges to C domain residues K273 and H406[37]. In TycB1 it was determined that the PCP interaction surface for the upstream C domain (TycA) is formed by four residues from helix αII - H44 (S -1), A50 (S $+5$), H56 (S $+11$), R57 (S $+12$)[142]. A helix αII residue, S526, was also

identified as the downstream C domain (TycB1) interaction motif of TycA-PCP[122].

Numerous studies to identify CP partner domain interaction residues and motifs have been carried out using the NRPS responsible for the biosynthesis of enterobactin in *E. coli*. This NRPS is composed of two modules; EntE (A) and EntB (ICL-ArCP) form module one, and EntF (C-A-PCP-Te) module two. All EntB-ArCP partner domain interactions occur *in trans* and all EntF-PCP partner domain interactions *in cis*. Two EntB-ArCP residues, D240 (S -5) and D244 (S -1), from the loop linking helices αI and αII and one residue, D263 (S $+18$) from helix αIII were determined to form the interaction motif for the A domain (EntE)[114]. The interface between EntB-ArCP and the downstream C domain (EntF) was shown to involve helix αII, M249 ($+$ 4), and helix αIII, F264 (S $+19$) and A268 ($+$ 23), residues[145]. In the SrfA-C multi-domain structure (C-A-PCP-Te) the PCP domain is in the A/H conformation and analysis of the inter-domain distances suggests the PCP domain is primed to interact with the upstream C domain. At the interface of these domains two residues M1007 (S $+4$) of helix αII and F1027 ($+$ 24) of helix αIII, previously identified in the EntB-ArCP EntF-C interface, form hydrophobic interactions with F24 and L28, and Y337 of the upstream C domain, respectively[38]. In the EntF-PCP domain two residues of helix αIII, G1027 and M1030, were identified as participating in the EntF-Te domain interface[144].

$\sim$80% of EntB-ArCP residues, from helix αI through to helix αIII, have been subjected to mutagenesis studies. This has revealed that most positions are tolerant to mutation - 36 of 44 residues examined showed low conservation - suggesting that only a few EntB-ArCP residues are involved in inter-domain recognition[143,145]. For a summary of CP residues identified as partner domain interaction residues see table 1.3. Figure 1.13 shows a number of these residues mapped onto the structures of TycC3 and EntB.

1.6 The Condensation Domain

As previously mentioned, the C domain is responsible for the elongation of the growing polypeptide chain. It serves to catalyse the condensation reaction between the downstream

Domain (colour Figure 1.12)	PCP	Active Serine	CP res	Helix/loop	Partner	Other[a]	Partner res	Ref
PPTase	TycC3 PCP †	S45	14 aa of αII	αII	AcpS	TE		122
(blue)	TycC3 PCP	S45	K47	αII	Sfp	TE	D40	142
	EntB ArCP ††	S245	D244R/A *	L αI-αII	EntD	TE		114
	EntB ArCP	S245	G242A, D244R **	L αI-αII	EntD, Sfp	TE		143
	EntF PCP	S1006	L1007	αII	EntD, Sfp	TE		144
	VibB ArCP →	S46	I47V	αII	Sfp	TE		146
	EntB ArCP	S245	V246					
A	HMWP2 ArCP →	S58	S49D, H66E	L αI-αII, αII	VibE	TUE		29
(orange)	VibB ArCP		E239, E256					
	EntB ArCP	S245			EntE	TUE	R437D, K473D	114
	EntB ArCP	S245	D240R, D263R	L αI-αII, αIII	EntE	TUE		114
			D244R *	L αI-αII				
	VibB ArCP →	S46	N38D, E70K	L αI-αII; αIII	EntE	TUE		146
	EntB ArCP	S245	N237, K269					
C	TycA PCP		S562A	αII	TycB1	TDE		122
(upstream - green)	HMWP2 ArCP →	S58	S49D, H66E	L αI-αII, αII	VibH	TDE		29
(downstream - red)	VibB ArCP		E239, E256					
	TycB1 PCP	S45⋆	H44, A50, H56, R57	αII	TycA	TUE		142
	EntB ArCP	S245	M249; F264, A268	αII; αIII	EntF	TDE		145
	VibB ArCP →	S46	E70K	αIII	EntF	TDE		146
	EntB ArCP	S245	K269					
	TycC5 PCP ‡	S43	R16	αI	TycC6	CDS	K273, H406	37
			E56	αII			K273, H406	
	SrfA-C PCP ‡‡	S1003⋆⋆	M1007	αII	SrfA-C	CUS	F24	38
			F10227	αIII			L28	
Te	EntF PCP	S1006	G1027A, M1030A	αIII	EntF	CDE		144
(cyan)								

Table 1.3: **PCP domain residues determined to interact with specific partner domains.** Where: Other[a] - T: trans, C: cis, U: upstream, D: downstream, E: Experiment (mutant), S: structurally characterised; → - mutated towards; † - TycC3-PCP pdb 1DNY and 2GDW (A/H), 2GDY (A), 2GDX (H); †† - EntB (ICL-ArCP) pdb 2FQ1; * - EntB ArCP D244R mutation showed only 35% modified with EntD (PPTase) and 13% WT activity with EntE the upstream A domain although PCP was a mixture of apo and holo; ** - EntB-ArCP D244A efficiently modified by EntD but not by Sfp; ⋆ - TycC3 numbering; ⋆⋆ - Active site serine mutated to alanine; ‡ - TycC PCP_5-C_6 pdb 2JGP; ‡‡ - SrfA-C (C-A-PCP-Te) pdb 2VSQ.

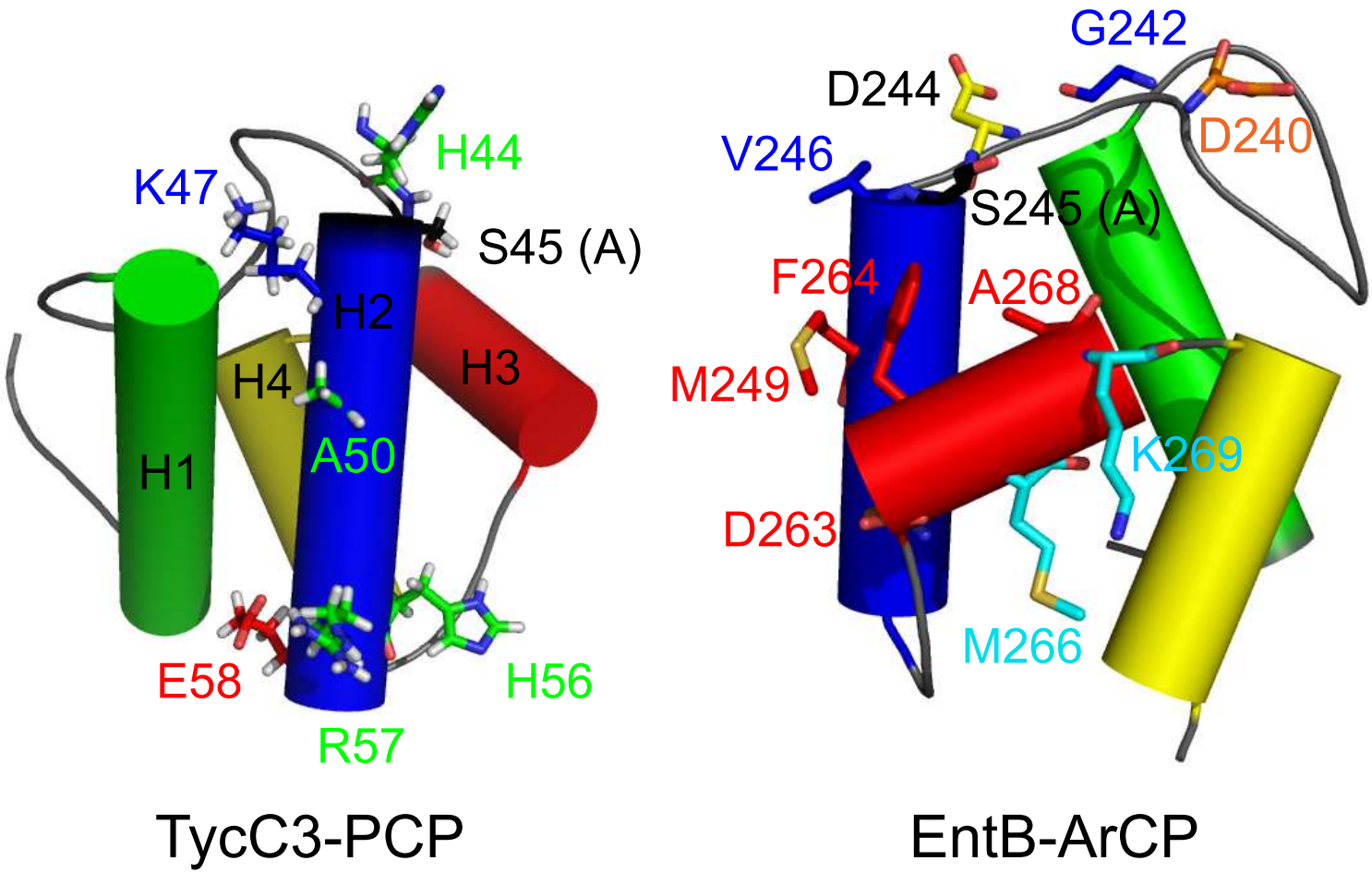


Figure 1.13: **The partner domain interaction residues of the NRPS CP domains TycC3 and EntB.** Helix α1 is coloured green, helix α2 in blue, helix α3 in red, helix α4 in yellow, the loops linking the helices (L1, L2 and L3) in grey and the active site serine residue in black. Residues that interact with PPTases are in blue, upstream C domains in green, downstream C domains in red, A domains in orange, Te domains in cyan, and those that interact with both PPTases and A domains are coloured yellow. Of the residues shown on TycC3-PCP H44, A50, H56 and R57 where identified from studies on TycB1-PCP[142] and residue E58 from studies on TycC5-PCP[37]. Of the residues shown on EntB-ArCP M266 and K269 were identified from studies on EntF-PCP (residues G1027 and M1030 respectively[144]. For ease of complete residue visibility TyC3-PCP is shown with helices α1 and α2 at the front and EntB-ArCP with α3 and α4 at the front.

PCP tethered amino acid and the peptidyl chain tethered via a Ppant arm to the upstream PCP domain. In 1996 Stein *et al.* proposed a revision to the multiple-carrier template model originally proposed by Lipmann in 1971[147]. In this model it was proposed that C domains contain an acceptor site (for the nucleophile) and a donor site (for the electrophile)[28]. In 2002, the determination of the X-ray crystal structure of VibH, a condensation enzyme from the *Vibrio cholerae* vibriobactin synthetase, and subsequent modelling studies reinforced this mode of action[148].

VibH (illustrated in figure 1.14) is a free-standing C domain. Unusually, one of its substrates is a small-molecule nucleophile, rather than a PCP-loaded amino acid. Multiple sequence alignments and secondary structure prediction have shown the similarity of VibH to other NRPS incorporated C domains, Cy and E domains. The structure of VibH is therefore representative of these three classes[148]. The mechanism by which C domains catalyse peptide bond formation and the proposed mechanism of action of the Cy domains is shown in figure 7.4 in appendix 7.1.2. C domains are structurally related to chloramphenicol acyltransferase (CAT) and dihydrolipoamide acyltransferases. The conserved motif, HHxxxDG, found in C and E domains is also present in CAT and E2p, the dihydrolipoamide acyltransferases of pyruvate dehydrogenase[117]. Comparison of VibH and the structures of CAT and E2p, reveals that C domains have a novel topology and therefore represent a new member of the CoA-dependent acyltransferase superfamily[148].

In VibH the motif H**H**xxxDG is located at the interface between the two domains. A solvent channel runs through the molecule between the two domains and allows access to the proposed catalytic His 126 residue (highlighted in bold in the motif) from either 'face' of the molecule[148].

Since structures of CAT have been determined with substrates CoA and chlorophenicol (CAM)[149,150] and E2p with substrates CoA and lipoamide[151–153], a comparison can be made between the location of the binding sites of these enzymes and potential binding sites in VibH. The location of the CoA, CAM and lipoamide substrates was mapped onto the N-terminal domain of VibH by superposition of the CAT complex and E2p complex monomer structures. In the resulting model, CoA would enter the solvent channel from the C-terminal

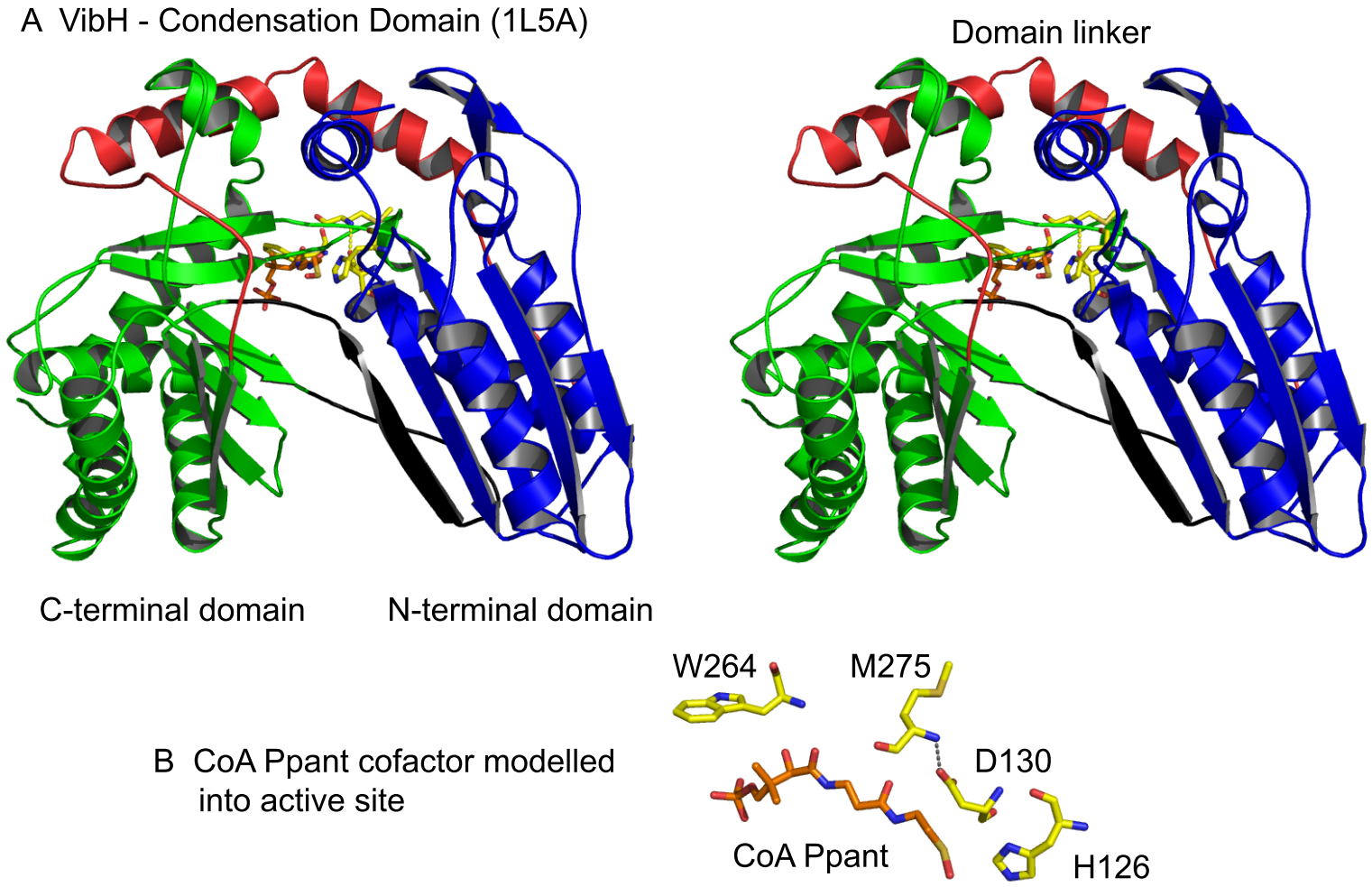


Figure 1.14: **Structure of the VibH monomer.** A)The N-terminal domain is in slate blue, the C-terminal domain in green, the region linking the domains in red, and the C-terminal domain β-strand that completes the N-terminal domain β-sheet in black. The CoA PPant moiety from CAT (pdb 3CLA [149] has been modelled into the VibH active site by superimposing the CAT monomer N-terminal domain onto the VibH N-terminal domain. B) residues H126, D130, W264 and M275 are shown as they are in close proximity to the PPant moiety.

face, positioning the CoA terminal thiol group in close proximity to His 126. By analogy, in VibH the 2,3-dihydrobenzoyl (DHB)-*S*-VibB substrate would bind to the C-terminal face of the protein with the Ppant arm extending into the solvent channel. Trp 264 is located at the C-terminal face surrounded by hydrophobic residues and projects into the solution. Experimental mutations of this residue have demonstrated its importance in the binding of VibB[148].

The CAT and E2p substrates, CAM and lipoamide respectively, map onto to the N-terminal face of the solvent channel and are positioned within 2.5–3 Å of the N_ϵ atom of His 126[148]. This model shows how both acyl donor and acceptor PCP substrates would be accommodated by the C domain. The PCP Ppant arms would enter from opposing open ends of the domain; the upstream PCP Ppant arm would enter the channel from the C-terminal face (the acceptor site) and the downstream PCP Ppant arm from the N-terminal face (the donor site)[148]. The C domain acceptor binding pocket exhibits strong stereoselectivity[31] and some selectivity towards the aminoacyl thioester[154]. These findings explain the role of the acceptor site in controlling the directionality and speed of chain elongation[34]. The donor site exhibits broader substrate specificity; yet once again the stereochemistry of the C-terminal amino acid of the peptidyl chain is important in substrate recognition[31].

These findings imply that the possibility of 'module swapping' (as discussed in section 1.3) as an NRPS engineering strategy, may be more feasible than previously thought and may possibly be more effective than the proposed strategy of A domain swapping. These theories may also prevent the implementation of a 'generic' A domain designed to exhibit relaxed substrate specificity[30].

1.7 The Epimerisation Domain

A substantial percentage of NRPs contain D-configurated amino acids. D-amino acid incorporation can be achieved by a D-amino acid selective A domain. This strategy is often employed in fungal systems, such as cyclosporin synthetase in *Tolypocladium niveum*[16]. In cyclosporin synthetase the D-Ala residue is produced by an external racemase[155]. An

E domain is, however, more commonly used to incorporate D-configurated amino acids[35]. These domains are usually located at the C-terminal end of the D-amino acid incorporating NRPS module. E domains promote the epimerisation of the C_α-carbon atom of the PCP-tethered amino acid or C-terminal amino acid of the growing polypeptide chain[156]. The proposed mechanism of action of E domains is shown in figure 7.5 of appendix 7.1.2. Since this reaction affords a D/L equilibrium, the downstream C domain is responsible for selecting only the D-enantiomer or diasteriomer[31]. E domains that function as insertions in A domains have been identified in HMWP2 of yersiniabactin synthetase and pyochelin synthetase E (pchE) from *Pseudomonas aeruginosa*. Unusually, the pchE E domain catalyses epimerisation after formation of the peptide bond[157].

Noncognate amino acids have been shown to be racemised by E domains but with lower efficiency[34]. Further studies showed aminoacyl-PCP could be epimerised by artificial E domains without a preceding C domain. Conversely, no epimerisation activity was shown when the aminoacyl-S-Ppant substrate was bound in a preceding cognate C domain. This suggests that the aminoacyl-PCP remains tightly bound in the C domain acceptor site until condensation occurs. A lower binding affinity is exhibited by the resulting peptidyl-PCP for the C domain acceptor site and it is then transferred to the subsequent E or C domain[158,159].

These observations provided key information regarding the timing and directionality of peptide elongation in NRPSs.

In the absence of an E domain crystal structure, sequence analysis has been performed that shows these domains exhibit similarity to C and Cy domains[148]. A conserved HHxxxDG motif, where the second histidine residue is presumed catalytic, is shared by C and E domains[156]. In E domains, this histidine residue is presumed to de-protonate and subsequently re-protonate the C_α carbon atom. Multiple sequence alignments of the E, C and Cy domains, together with a comparison of the actual (VibH) and predicted secondary structures for these domains have been performed. These have demonstrated the presence of a 13 residue insertion in E domains at the C-terminal end of the proposed solvent channel and active site. This indicates the C-terminal face of the E domain may be blocked so only the N-terminal face may be responsible for PCP binding[148].

1.8 The Thioesterase Domain

The Te domain terminates the elongation process, releasing the final peptide product from the NRPS into solution. This occurs in a two-step mechanism. The first step is the formation of an acyl-*O*-Te intermediate. During the second step this intermediate is attacked by *either* a water molecule (hydrolysis)[32] to produce a linear peptide, or by an intramolecular nucleophile (macrocyclic release)[160] which results in a macrocyclic product. The later strategy is more commonly observed, possibly due to the resistance of these cyclic structures to proteolytic breakdown[161]. The mechanism of chain release by the Te domain is illustrated in figure 7.6 in appendix 7.1.2.

Prior to the release reaction, individual Te domains can also catalyse other highly specialised reactions which can increase the biological activity of NRPs. Examples of such reactions include oligomerisation of the peptide chain as observed in Gramicidin S and enterobactin. Gramicidin S comprises two identical peptides that are bridged head to tail. A complete *in vitro* characterisation of the excised iterative Te domain responsible for the oligomerisation of Gramicidin S was recently described by Hoyer *et al.*[162]. Enterobactin is a trimer that has been cyclised to produce a macrolactone. The Te domain in the NRPS that produces cyclosporine A is responsible for amide bond formation between the N-terminal amino group and the C terminus of the peptide. In surfactin, the lactonisation of the C-terminal carboxyl group thioester of the peptide and the hydroxyl group of an N-terminal β-hydroxyl fatty acid is catalysed by the Te domain. Te domains can also produce branched cyclic structures.

Much of the versatility of NRP biological activity is attributed to the ability of the Te domains to catalyse such wide ranging reactions. This additional degree of Te domain specialisation requires great diversification of the Te domains which is, in turn, reflected in the low sequence identity between members of this family[6,163]. Te domain catalysed reactions are similar to those catalysed by serine esterases and lipases. The affiliation of Te domains with this group of α/β-hydrolases was confirmed when the structure of the lipopeptide antibiotic surfactin Te domain (SrfTe) was determined[163].

The structure of the Te domain (SrfTe) from the C-terminal surfactin synthetase (SrfA-C subunit) from *Bacillus subtilis* JH642 was determined by Bruner *et al.* in 2002[163]. This Te domain consists of the last 235 residues in the SrfA-C subunit. SrfTe is a globular protein with a pronounced bowl-like cavity that harbours the active site. The SrfTE structure displays the characteristic fold of the α/β-hydrolases. It differs from the generalised α/β-hydrolase secondary structure model in that strand β1 and helix αD are missing, and αE is a greatly reduced 3_{10} helix. α/β-hydrolases commonly contain a catalytic triad. In SrfTE, this catalytic triad comprises residues Ser 80, His 207 and Asp 107. These residues are responsible for the formation of the peptidyl-*O*-Te intermediate[163].

The crystal structure (pdb 1JMK) contains two independent monomers: an 'open' state monomer and a 'closed' state monomer, as illustrated in figure 1.15. The two structures differ between sheet β6 and helix αA_3, where three helices (αL_1, α L_2 and α L_3) form a 'lid' which reaches over the active site. The 'open' and 'closed' state structures of SrfTe are shown alongside the structure of the Te domain from FenB, FenBTe, in figure 1.15. In the structure of FenBTe[37] the residues that would form the flexible region immediately preceding the lid structure (T123–S129) are missing. The lid region of FenBTe is 12 residues shorter than that of SrfTe - lacking the region corresponding to helix Lα1 - and forms a long 16 Å helical segment, αL-helix, from residues L132 to K143 that protrudes markedly from the Te domain[37].

In the 'open' monomer, seen in figure 1.15.A., the lid is folded away from the active site to allow unrestricted access. The formation of a short antiparallel β sheet (residues 110–112 and 187–188) accompanies this movement. In the 'closed' monomer structure (figure 1.15.B.) the 'lid' obstructs the active site. No determinable electron density was observed for residues 116 to 123 in the 'closed' structure, suggesting flexibility in this region[163]. To characterise the active site further, SrfTE was cocrystallised with an analog of the *N*-acyl-heptapeptidyl-*N*-acetylcysteamine thioester substrate, *N*-acyl-heptapeptidyl-SNAC minus the β-hydroxy group on the fatty chain. This analog allows for the Te acylation stage of the reaction but not the cyclising deacylation[163]. SrfTE has also been co-crystallised with boronate inhibitors[164]. This led to the identification of a hydrophobic binding pocket for

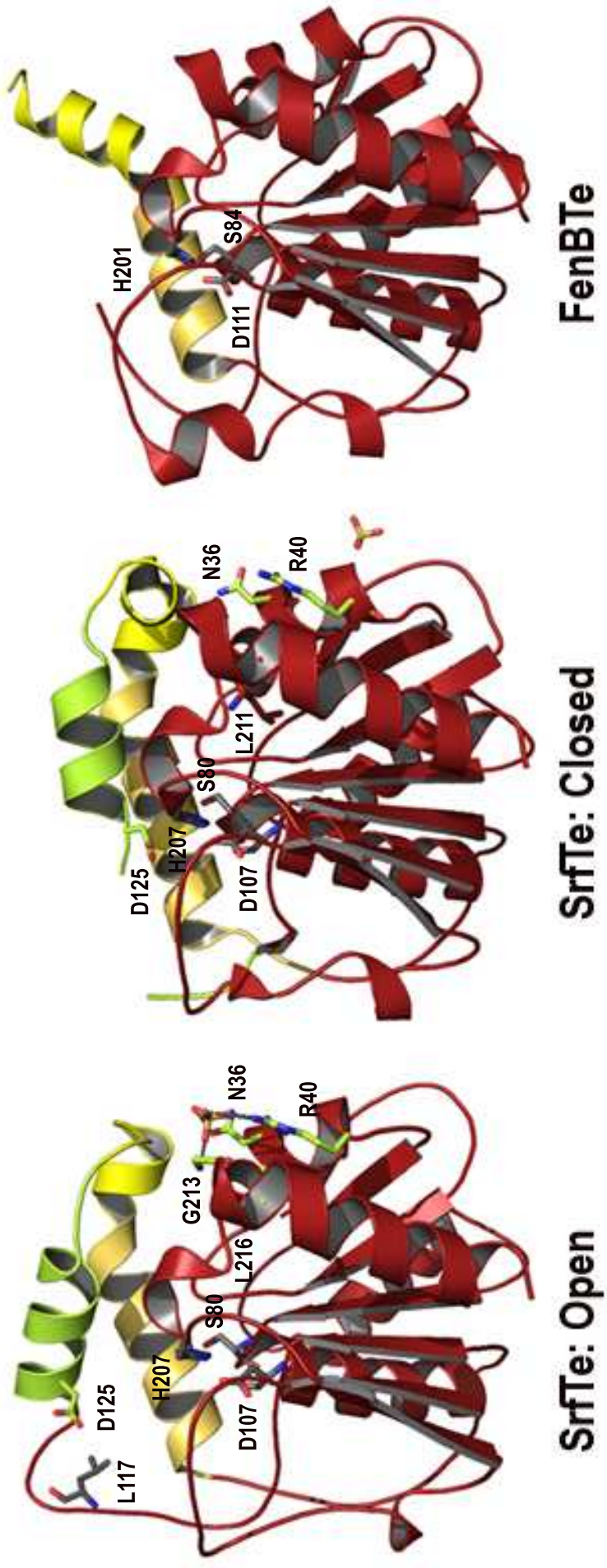


Figure 1.15: **Structures of the SrfA-C and FenB Te domains.** SrfTe is shown with both the 'open' (left) and 'closed' conformations of the active site. The lid region (αL_1, αL_2 and αL_3) is in yellow. Residues S80, D107, and H207 form the catalytic triad. In the open conformation a sulphate ion - presumed to mimic the binding position of the PPant arm - is coordinated by residues N36, R40 and G213. Residues L117 in the open conformation and D125 in both conformations are highlighted as these two residues are thought to restrict access to the Leu pocket in the closed conformation[163]. The FenBTe catalytic triad is formed by S84, D111 and H201.

the two C-terminal amino acids (D-Leu$_6$ and D-Leu$_7$) of the substrate[163,164].

No determinable electron density was obtained for the rest of the peptidyl chain. Modelling of the substrate in the Te domain active site places the substrate in a hydrophobic cavity with the peptidyl N-terminus lying adjacent to the active site serine residue[163]. NADPH-dependent Red domains catalyse the reduction of heterocyclic by the addition of two electrons. Such a reaction is seen in one of the rings of yersiniabactin and pyochelin, where thiazoline is reduced into thiazolidine[165]. In addition to this, Red domains can also replace Te domains to catalyse peptide release.

1.9 Additional Tailoring Domains

N-Mt and C-Mt domains are responsible for the N- or C- methylation of amino acid residues respectively. This methylation makes the peptide less susceptible to proteolytic breakdown[161]. *S*-adenosyl methionine (SAM) is used as a methyl donor by both types of Mt domains. The mechanism of methylation catalysed by N-Mt domains is illustrated in figure 7.5 in appendix 7.1.2.

Oxidation domains are proteins composed of $\sim$250 amino acids. Ox and Red domains can catalyse the change of oxidation state in oxazoline and thiazoline rings. Bleomycin[166], myxothiazol[167] and epothilone[168] synthetases all possess Ox domains in at least one of their modules. Myxothiazol contains two Ox domains. The in-frame deletion of one of these domains does not abolish the production of a thiazole final product. The other Ox domain in the synthetase is thought to oxidise both thiazolines[169]. The N-terminal peptide end can be modified by N-formylation as in anabaenopeptilide 90-A and linear gramicidin A. This modification is catalysed by a F domain[25,78].

1.10 Linear NRPSs

Linear (type A) NRPSs include the bacitracin, surfactin, tyrocidine, cyclosporin, pristinamycin, fengycin and complestatin synthetases. In linear NRPSs an initiation module is followed by an elongation module and the peptide chain is released from the enzyme using a Te domain. In fungal NRPSs this Te domain is often replaced by a specialised C domain which catalyses such cyclic reactions. The number and order of the modules within a linear NRPS determines the sequence of the peptide product. The domain organisation in a linear NRPS can be summarised as A-PCP-(C-A-PCP) $_{n-1}$-Te where n is both the number of NRPS modules and the number of amino acids combined into the product[5].

1.11 Aims of Thesis

As antibiotic resistance is increasing more rapidly than new antibiotics are produced and/or discovered, there is an increasing need to determine new ways of designing novel antibiotics. A potential avenue for this is the exploitation of NRPSs. As the primary, yet not exclusive, determinant of substrate selectivity in NRPSs, the A domain is a target for genetic manipulation to alter the amino acids incorporated by an NRPS module. In order to do so however, a detailed molecular understanding of the A domains is required. While A domains have been studied extensively, knowledge of the selectivity mechanism and domain dynamics is still relatively rudimentary. Alternation between two conformations has been proposed as a strategy used by members of the adenylate-forming superfamily to reconfigure a single active site to perform two different reactions.

The work presented in this thesis was initiated in March 2005 (and completed in December 2006). At this time while an acetyl CoA synthetase structure (from the wider ANL superfamily) had been determined in an alternate conformation (now referred to as conformation 2), all of the available A domain structures were in conformation 1, and it was not believed that the A domains utilised domain rotation or alternation for catalysis. In 2008 structures of chorobenzoate CoA ligase were determined with the substrates from both half reactions

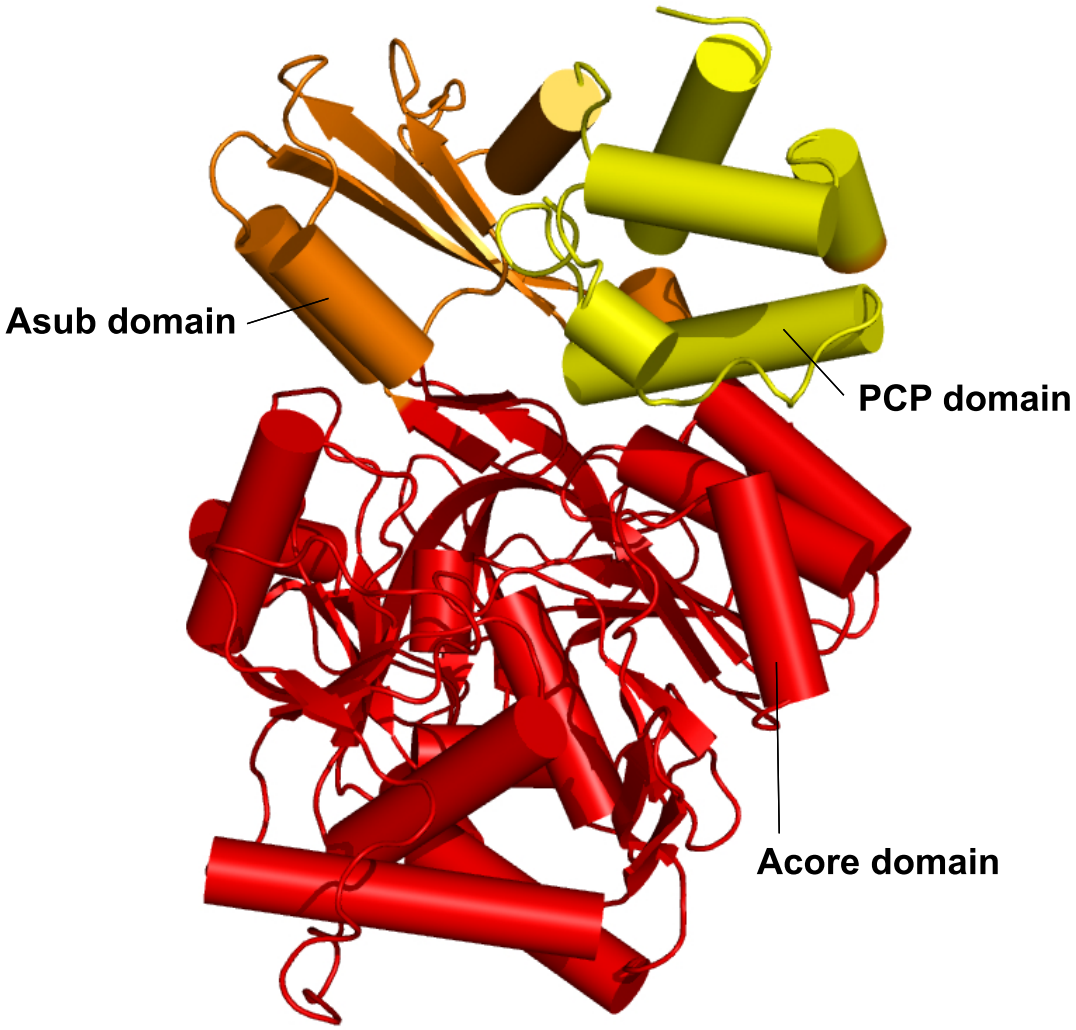


Figure 1.16: **Structure of PA1221 A domain and PCP domain.** The A_{core} domain is shown in red and the A_{sub} domain in orange. The PCP domain is shown in yellow.

in two different conformations, providing evidence that two distinct conformations were utilised for catalysis[53]. Recently Mitchell *et al*[170] determined the structure of PA1221, a novel NRPS A domain and the associated PCP domain in the second half conformation, figure 1.16, providing evidence that the A domains of NRPSs also utilise two distinct conformations of the same protein to catalyse the two half reactions.

One way to probe conformation and examine the interaction between proteins and ligands is by using computer simulation, especially Molecular Dynamics (MD). This can provide information at the molecular level that is complementary to experiment and can, therefore, further our understanding of a system. To date no molecular simulation study of the A domains has been reported in the literature.

The MD simulation studies presented in this thesis were designed to:

1. Explore the dynamics of the PheA A domain in the presence and absence of the hydrolysed products of the first half reaction;

2. Explore the dynamical behaviour of PheA in the presence of first half reaction noncognate substrates;

3. Use molecular modelling techniques (homology modelling, docking and MD) to study an A domain from an iterative NRPS.

Chapter 2

Computational Methods

2.1 Secondary Structure Prediction

It has been known for 40 years that the information required for a protein to adopt its native fold is contained within its primary sequence [171,172] with some contribution from its native solution environment. Although it is know that chaperones are required for folding in some instances, a large amount of research supports the hypothesis that the native conformation for most proteins corresponds to the lowest free energy conformation of that sequence. This suggests it should be possible to determine the fold from the amino acid sequence. Limited computing resources and inaccuracies in experimentally determining the basic parameters prevent successful prediction of protein structure from first principles [173]. As a result the most commonly used and successful prediction methods are knowledge-based. Whilst the field has expanded rapidly in the last 15 years, it is still not possible to predict accurately the fold of a protein from its sequence; rather, emphasis is placed on correctly predicting the basic characteristics and elements of the fold. Predicting the location of helices and sheets is the first stage in determining the fold.

Methods to predict secondary structure have been around for over a quarter of a century. First generation prediction methods developed in the 1960s and 1970s, exploited two aspects of protein structure: the intrinsic secondary structure propensities of amino acids and the hydropathic nature of amino acids. These methods, the most famous of which are the Chou and Fasman method from 1974 [174] and the Garnier, Osguthorpe and Robson (GOR) method from 1978 [175], were originally reported to be between 70 and 80% accurate, however, they have been shown to be only 56-60% accurate [176].

The fact that specific segments of conserved residues exhibit preference for specific secondary structural elements has lead to the second-generation of prediction methods [176]. Methods compiling propensities for stretches of 3 - 51 adjacent residues, took into account the local environment of individual residues. These methods redefined the problem as essentially one of pattern classification and therefore recognition. These initial prediction methods were based on statistical analyses performed on available proteins to determine the propensity of each individual amino acid to partake in helices or sheets. Many differ-

ent theoretical techniques were applied to rationalise the relationships in these segments of residues, including graph theory, linear and multilinear statistics, nearest neighbour algorithms, molecular dynamics and neural networks.

In the early 1990s the level of prediction accuracy had reached an apparent ceiling of approximately 60%. Third-generation methods, initiated by Rost and co-workers in the PHDSec program, use a particular combination of neural networks and evolutionary information. Sequence profiles based on aligned families of proteins are used to train neural networks. Prior to this, neural networks employed to locate regions of secondary structure were trained solely on binary encoded single amino acid sequences. Although often presented with binary values, neural networks can be trained with arbitrary real values. In these profile based methods, the training inputs to the network are typically the probabilities of occurrence for the 20 amino acids. Using these profiles, the prediction accuracy can be raised to as much as 70-77% [177].

The combination of neural networks and evolutionary information solved two critical problems in secondary structure prediction and raised prediction accuracy to above 70% [178,179]. The length of elements, previously underestimated by as much as half, were more accurately predicted. In addition the strand location accuracy was significantly improved, as previously this was not much more precise than random predictions. The incorporation of evolutionary information aided more accurate predictions. 67% of residues within a protein can be exchanged without any significant alterations to the protein structure [180,181]. However it is also true that just a few mutations in a critical region can destabilise a protein. Evolution within protein families can generate maximal diversity by exploiting mutations of structurally non critical residues. Multiple alignments of protein families can provide residue exchange patterns that yield information regarding structural element location and profiles obtained from these alignments can provide nonlocal three dimensional data. Family profiles are used in all of the current most accurate prediction methods [181].

Although the PHDSec program was the first to surpass the 70% threshold for accuracy it is no longer the most accurate prediction method. PSIPRED [177] implements automated, iterative PSI-BLAST searches to produce profiles which are not tainted by unrelated proteins. It

also uses a neural network similar to that of PHD. One of the most accurate more recently developed methods is SSpro. This method uses profiles and neural networks in combination with an improved algorithm.

2.2 Homology Modelling

In order to realise the full potential of the genome sequencing projects, the function of proteins encoded by the genomes need to be assigned, understood, controlled and modified. This is largely facilitated by knowledge of the native three-dimensional protein structure. Currently the structures of only a fraction of known protein sequences have been experimentally determined. Many biomolecules are not suited to structure determination using current experimental techniques, due to size or environment restrictions, e.g. membrane proteins. For these reasons protein modeling is one of the most rapidly expanding and debated areas of structural bioinformatics.

The prediction of the three dimensional structure of a protein from its one dimension sequence is the subject of the field of protein modelling. The goal of this field is to be able to predict the structure with an accuracy that is comparable with experimentally determined results. Reasonable applications of any theoretical model depend on its accuracy. Errors rarely occur in the functionally important regions in theoretical models as such regions, which include the active site, tend to be more highly conserved in evolution than the rest of the fold. Lower quality models can therefore be used to analyse ligand binding interactions or predict a likely ligand from the cleft volume [182–185].

The three main approaches to predicting the three-dimensional structures of proteins are homology or comparative modelling, fold recognition and first principles or ab initio techniques. The most accurate models are generally obtained by homology modelling for two main reasons. Firstly because the structure of a protein is uniquely determined by its primary sequence, knowing the sequence should therefore be sufficient to obtain the structure. Secondly, the structure of a protein is more stable and changes more slowly than the sequence during evolution. An accurate limit for the structure sequence relationship was

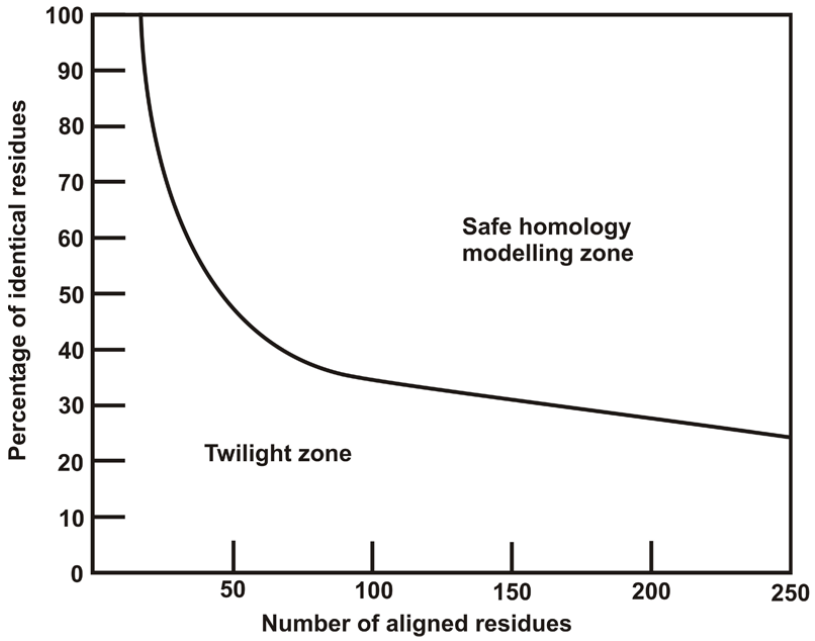


Figure 2.1: **The two zones of sequence alignments.** Image adapted from [185]
.

developed due to the exponential growth of the Protein Data Bank (PDB). Two sequences are practically guaranteed to adopt a similar structure as long as the length of the sequences and the percentage of identical residues fall in the region designated as 'safe', see figure 2.1 [185].

Homology modelling consists of four steps; template identification, alignment of the template and target sequences, building the model and assessment of the structure. The template is the structure on which the target sequence is modelled. More than one template can be used and is often required when regions of one template have not been experimentally determined or are not highly conserved. Central to this approach of protein modelling is the observation that indels are usually accommodated by small local conformational changes which can only occur in structurally variable regions, usually loops or turns, while the structurally conserved regions remain largely unchanged [186].

2.2.1 Template Identification

The starting point in comparative modelling is to identify a suitable template sequence. This can be done using similarity searching programs such as BLAST [187] or FASTA [188],

using PSI-BLAST to detect more distant homologs or by threading or fold recognition techniques. Considerations when choosing an appropriate template include the family or subfamily to which the target belongs, the environment, e.g. inclusion of cofactors, in which the template structure was determined and the quality of the experimental template structure. The priority of the criteria for template selection depends on the purpose for which the model will be used. If a protein-ligand model is needed, choosing a template which contains an equivalent ligand may be more important than the template structure resolution [189].

2.2.2 Target-Template Alignment and Refinement

Once the template has been selected a specialised method should be used to align the sequences. As the sequence identity between two sequences decreases the number of indels increases and obtaining the optimal target-template alignment, particularly using automated techniques, becomes more difficult. For sequences with identity over 40% the alignment is almost always correct. Sequences that share between 30 and 40% idenitity often contain regions of low sequence similarity. When the sequence identity falls below 30% and into the 'twilight zone', obtaining a good alignment becomes more difficult. At this level the alignment contains more indels, and therefore more errors. At 30% identity only 20% of residues are likely to be correctly aligned [189].

As a result, the accuracy of comparative modelling is dependent on the level of sequence identity between the target and template sequences. High accuracy homology models, those with an r.m.s. error comparable to that of medium-resolution NMR structures (~ 1 $\dot{A}$), are obtained when there is $\geq$ 50% identity between template and target. Medium accuracy models, defined as those with an r.m.s. error of 1.5 $\dot{A}$ or less for $\sim$ 90% of the main chain atoms, can be expected from alignments between sequences that share 30 - 50% identity. At less then 30% seqeunce identity, low accuracy models are produced [183].

No comparative modelling method can recover from an incorrect alignment. Therefore, effort must be invested in generating an accurate alignment. When sequence identity is low,

sequence information and multiple structures can be relied upon to guide the alignment. Information gained from secondary structure prediction tools can be used to ensure that gaps are not placed in; secondary structure elements, residues that are far apart in space or buried regions. Any invariant or highly conserved regions of sequence can also be used to guide the alignment. As the alignment is such a crucial stage in homology modelling it is wise to manually inspect and edit the alignment with reference to the template structure. Model evaluation is more reliable than alignment evaluation. If the choice of alignment is unclear, generating three-dimensional models for all alternative alignments and evaluating the corresponding model may be fruitful [189].

As part of some earlier sequence analysis work, not presented in this thesis, a profile hidden markov model (pHMM) was constructed using an A domain multiple sequence alignment which was built using ClustalX and manually refined guided by secondary structure predictions. This pHMM was used to guide and assess the alignments used as input for the homology modelling work presented in Chapter 5.

2.2.3 Model Building

A variety of methods are available to construct the model protein from the alignment. These methods can be divided into three groups: rigid-body assembly (traditional homology modelling), segment matching and modelling by satisfaction of spatial restraints [189]. Rigid-body assembly is the original method and is still popular today. The MODELLER program written by Andrej Sali [190], which utilises spatial restraints, was used to generate the homology model presented in this thesis and therefore will be discussed alongside the more traditional rigid-body approach.

Modeller was selected for the homology modelling work presented in this thesis as at the time it was consistently assessed as a high performing modelling tool when models are evaluated for physiochemical correctness and structural similarity both in research studies[191] and at the biennial Critical Assessment of protein Structure Prediction (CASP) experiment.

2.2.4 Rigid-Body Assembly

The rigid-body assembly method uses a small number of rigid bodies, obtained from aligned protein structures, to assemble a protein model. The technique utilises the conserved core, variable loop and side chain regions and protein folds. The COMPOSER [192] program performs this dissection automatically. After selection, the template structures are superimposed to allow calculation of the model frame-work by averaging over the C_α atoms of the structurally conserved regions. The model frame-work is built up bit by bit, generating the target molecule core-region main chain atoms by superposing core segments from templates with the highest sequence identity in comparative regions. Database searching is utilised to construct the variable loop regions. Loops with a comparable sequence and which fit the anchor regions are selected. Side chains are modelled based on equivalent residues in the template structure and their intrinsic conformational preferences. Refinement of the model is achieved using restrained energy minimisation or Molecular Dynamics (MD) [189].

2.2.5 Satisfaction of Spatial Restraints

MODELLER is one of the most successful programs for generating protein homology models. The program works from an alignment, but the model is built in one step using a combination of conjugate gradients and molecular dynamics (MD) with simulated annealing. A conventional MD force field, CHARMM22[193], is used to enforce proper stereochemistry. Additional spatial restraints, a set of optimal inter-Cα distances and dihedral angles calculated from the template structure(s), are embedded into the MD force field. This restraint-embedded force field is termed a molecular probability density function (PDF). In the case where more than one template structure is available the PDF is weighted so that more highly conserved regions have stronger restraints than those which vary in structure. The resulting model(s) satisfy the spatial restraints as well as possible. Additionally, loop regions can be independently remodeled and refined, the level of overall refinement can be controlled and varied, additional experimentally derived restraints can be added to guide the modelling and for each alignment a number of different models may be produced [189,190].

2.2.6 Loop Modelling

Changes in loop conformation are notoriously difficult to predict. Even in the absence of indels, loop conformations can vary widely between template and target structures. Reasons for this include, the steric hindrance of residues in close proximity to the loop altering its position, the involvement of surface loops in crystal contacts and the residue composition of the loop affecting the conformations it may physically achieve. The two main approaches to loop modelling are knowledge based and energy based. The knowledge based approach involves interrogating the PDB to locate loops with endpoints that correspond to those between which the loop has to be inserted and then copies the relevant loop conformation, as in the COMPOSER program. The energy based method uses an energy function to judge the quality of the loop, as in ab initio fold prediction. The function is then optimised using either MD or Monte Carlo (MC) to generate the more favourable loop conformation [185]. These methods have a reasonable chance of predicting a loop conformation that superimposes well on the true structure for shorts loops of five to eight residues.

2.2.7 Side-chain Modelling

Methods for generating side-chain conformations in traditional homology modelling approaches are usually at least partially knowledge based. Libraries of common rotamers extracted from high resolution X-ray structures are used. Available rotamers for each required residue are tried and each one scored using an energy function. The choice of neighbouring rotamer is affected by the previous rotamer choice, therefore producing a combinatorial explosion which is computationally demanding. The search space is greatly reduced as certain backbone conformations strongly favour specific rotamers [185].

2.2.8 Model Optimisation, Validation and Assessment

Energy minimisation can be used to optimize and refine the model, but it must be used carefully. Often it is not appropriate to minimise the complete structure as this requires

enormous accuracy in the energy function. The use of MD to refine structures generated by comparative modelling is still a matter of debate. When the model itself is close to the 'true' structure MD can prove fruitful, however a 'bad' model can be made worse [194]. Both these methods also rely on the accuracy of the chosen force field.

Ways of assessing the quality of a model include examining its geometry, stereochemistry, and other structural properties. Systematic analysis of existing structures within the PDB has provided a knowledge base of what is considered normal for proteins. Comparisons between this wealth of structural data and a model structure can be used to assess errors and quality. General checks, ideal bond lengths or bond and torsion angles, are less suitable for theoretical models than experimentally determined structures, as modern comparative modelling programs rarely make such errors. The most powerful check of steroechemical quality is the Ramachandran plot [195]. For every amino acid in the protein, excluding the two terminal residues, the ψ main-chain torsion angles are plotted against the ϕ main-chain torsion angles. In the resulting plot, favourable regions and regions which, due to the steric hindrance of the side-chain atoms, are unfavourable become apparent from the clustering of the points. High quality experimentally determined protein structures typically have well over $>90\%$ of their residues in the most favourable regions. Other parameters that can be assessed to validate protein structures include side chain torsion angles, the number of bad and unfavourable atom-atom contacts, inside/outside distributions of apolar/polar residues and the number of unsatisfied hydrogen bond donors [196]. Such checks can be performed using the PROCHECK program [197,198].

Root Mean Square Deviation (RMSD) is a conventional measure of model quality. It can be used to assess how similar the model structure is to the template, or any other structure. Structures of the same protein solved either by different groups or under different conditions typically exhibit a C-α RMSD of $\sim 0.6\ \dot{A}$. Values in this range are therefore the benchmark of homology modelling. The RMSD value effectively represents the difference between two complete sets of coordinates, something that is hard to represent in a single value. There are therefore some problems with this method of assessment. For example a region of locally correct structure interspaced with a short run of atoms that are badly placed would not be

truly reflected in a RMSD score [182]. The Z-score can offer another way of statistically representing the similarities or differences between a pair of protein structures.

2.2.9 Statistical Significance

As homologous sequences from different organisms can have a sequence identity as low as 25%, often another method is required to assess the significance of an alignment. This can be done by comparing the score of the proposed alignment with the scores from the alignment of the query sequence with other randomly chosen sequences. This value determines whether the proposed alignment is better than one selected at random. The mean and standard deviation of the random alignment scores can be obtained to determine whether the score of the original alignment is particularly high. The Z-score, equation 2.1, is a measure of whether the alignment is an outlier from the population:

$$\mathrm{Z-score} = \frac{\text{score - mean}}{\text{standard deviation}} \tag{2.1}$$

If the alignment is not better than average random permutations of the sequence, then it may have arisen by chance and a Z-score of zero is obtained. As the value of the Z-score increases, the likelihood of the alignment occurring by chance decreases. Commonly, Z-scores greater than or equal to 5 are taken to be significant [199].

The PROSA II program calculates atoms radial distribution functions and converts them into an energy-like quantity. Using this method, misfolded structures and incorrect folds can be identified [200]. Discrete Optimized Protein Energy (DOPE) is a similar method for scoring protein structures. Using probability theory, an atomic distance-dependent statistical potential, that does not depend on any adjustable parameters, is derived from a sample of native structures. An energy profile is produced that can be used to identify misfolded regions of structure [201].

2.3 Docking Methods

Molecular docking methods attempt to predict the structure(s) of the intermolecular complex between two or more molecules by sampling conformations of the ligand in reference to the active site of the protein. As the majority of docking algorithms produce a large number of possible structures, scoring functions are used to evaluate which ligand conformation best complements the protein binding site [202].

Docking can be performed manually and can be very effective if the expected binding mode is known; if for example the binding mode of a closely related ligand has been determined. It should be noted however that X-ray crystallographic experiments have shown that very similar inhibitors can adopt distinctly different binding modes [203].

In addition to the conformational degrees of freedom of each individual molecule, one molecule (e.g. the ligand) has six degrees of translational and rotational freedom to another (e.g. the protein). Docking algorithms can be characterised based on the number of degrees of freedom that they disregard and the currently available techniques can be broadly classified into two groups; matching methods and docking simulation methods [203].

2.3.1 Matching Methods

Matching methods use the simple approximation of treating each molecule as a rigid entity and exploring only the six degrees of translational and rotational freedom between them. These rigid-body approximation methods identify ligands with a high degree of shape complementarity to the binding site. They work by creating a model of the active site into which the ligand is docked by matching the geometry of the ligand to that of the binding pocket [203].

The DOCK program of Kuntz and co-workers [204] is a matching method docking program which uses a simple function containing only two terms, hard sphere repulsions and hydrogen bonding, to treat the geometric interaction of two rigid bodies. In this method the binding site (pockets or grooves) of the protein (receptor) is represented as a series of over-

lapping spheres of varying radii which fill the active site. Each sphere touches the binding site surface at only two points. The ligand molecule is also represented by a set of spheres. The ligand spheres are then matched to the spheres representing the binding pocket and matching sets identified in which the distances between all atom spheres in a set must correspond, within a given tolerance, to the internal distances between the equivalent binding pocket spheres. This method for indentifying geometrically similar clusters of spheres in the ligand and binding pocket sphere sets negates the need to perform explicit rotations of one structure relative to the other. A least squares fit of the atom spheres to the binding site sphere centres is then performed to orient the ligand within the binding site, reducing atom overlaps and ensuring hydrogen-bonding partners. The ligand orientation is then checked for unfavourable steric interactions with the protein. The interaction energy is computed for acceptable ligand orientations and used to score the binding mode. Different ligand orientations are generated by matching alternate sets of atoms and sphere centres [204].

2.3.2 Docking Simulation Methods - Flexible Ligand Search Algorithms

Docking simulation methods allow modelling of the flexibility within the ligand. Such methods are more physically detailed and typically slower than matching techniques. The majority of these methods treat the protein as a rigid body only considering the conformational space of the ligand. Docking approaches that consider both ligand and protein flexibility have been developed in recent years and progress within the field is discussed in a recent review by B-Rao and co-workers [205].

The methods that model flexibility within the ligand can be broadly categorised into three types: systematic, random or stochastic, and classical simulation.

Systematic docking algorithms attempt to explore all possible degrees of freedom within a ligand. Conformational search, fragmentation and database are all methods utilised in systematic docking algorithms. Conformational search methods systematically rotate all rotatable bonds in the ligand through $360\,^{\circ}$ using a fixed increment, generating and subsequently evaluating all possible combinations. Application of this method is very limited

as the number of different generated structures increases with an increase in the number of rotatable bonds; this is an example of a combinatorial explosion. Dimensionality of the problem can be reduced by applying constraints and restraints to the ligand [206]. In fragmentation methods the ligand is incrementally grown into clefts (binding pocket) of a protein structure either by docking in a rigid fragment of the ligand and then successively adding flexible regions of the ligand, or by docking several fragments into the binding pocket and then linking them covalently[206]. Examples of programs that have implemented a fragmentation search method include LUDI [207], ADAM [208], FlexX [209] and DOCK [210].

Database methods address the issue of combinatorial explosion introduced as a result of considering ligand flexibility by using libraries of ligand conformations. The program FLOG uses distance geometry constraints to generate a library of 25 database ligand conformations that are subsequently docked into the rigid protein [211].

Random or stochastic algorithms primarily use either a Monte Carlo (MC) algorithm or genetic algorithm (GA). These methods make incremental changes to the ligand or population of ligands, which are either accepted or rejected at each step dependent on a predetermined probability function.

In MC methods a Boltzmann probability forms the basis of the criteria upon which each new ligand is evaluated[206]. Programs that use MC-based algorithms include DockVision [212,213], Prodock [214], and MCDOCK [215].

Genetic algorithms [216] pioneered by John Holland are global heuristic search algorithms inspired by evolutionary biology. GAs randomly generate an initial population of candidate solutions (individuals), which are represented abstractly (commonly by binary strings) by genes organised into a chromosome. The individual solutions within a population evolve over generations towards better solutions. Only a proportion of the existing population is carried forward to produce the new generation. The *selection* of individuals to progress is dependent on *fitness*. Pairs of selected individuals (parents) combine, a process termed *crossover*, to produce *offspring* which then form the new population. Some of these offspring additionally undergo random changes or *mutations*. New iterations of the algorithm

begin with each new generation and the process only terminates when either a solution with a satisfactory fitness level has been generated, or the predetermined maximum number of iterations has been reached. If the latter is true a solution of satisfactory fitness level may not have been obtained [203]. Examples of programs that include a GA method for molecular docking are; GOLD [217,218], DIVALI [219], AutoDock version 3 [220], and DARWIN [221].

Simulation methods applied to molecular docking include simulated annealing and energy minimization methods. Application of simulated annealing to the docking problem circumvents the limitations of MD for crossing high-energy barriers in the energy landscape of a biological system and allow for a search of greater conformational space. Energy minimization methods are rarely used as a docking search technique, they are however, commonly used to optimize potential ligand conformations [206].

2.3.3 AutoDock

In the AutoDock program of Morris and co-workers the docking simulation can be carried out with one of the following search methods: MC simulated annealing; GA; local search (LS); global-local search method, the Lamarckian genetic algorithm (LGA).

Versions 1 and 2 of the Autodock program contained only a MC simulated annealing search option called the Metropolis method. The simulated annealing technique has both global and local search attributes; performing a global search at higher temperatures and a more localised search at lower temperatures. This method had limitations however when attempting to dock ligands with more than eight rotatable bonds. The GA, LS and hybrid LGA method, developed to address the limitations of the *Metropolis method*, were introduced in version 3 of the program. Version 3 also contains an empirical binding free energy force field which was also developed to allow the prediction of binding free energies of docked ligands with greater accuracy [220].

Atomic affinity potentials pre-calculated for each atom type in the ligand enable rapid energy evaluation. The AutoGrid routine embeds the protein in a three-dimensional grid placing a probe at each grid point. The energy of interaction of this single atom with the protein

is assigned to the grid point. An affinity grid is calculated for each type of atom in the ligand - typically carbon, oxygen, nitrogen and hydrogen - in addition to a grid of electrostatic potential, using either a Poisson-Boltzmann finite difference method or a point charge of $+1$. Tri-linear interpolation of affinity values of the eight grid points surrounding each ligand atom is used to determine the energetics of a particular ligand configuration. The electrostatic interaction is obtained by interpolating the values of the electrostatic potential and multiplying by the charge on the atom. These grids mean the time to perform the energy calculation is proportional only to the number of atoms in the ligand and independent of the number of atoms in the protein [222].

Genetic Algorithms in Molecular Docking

In molecular docking the position of the ligand with respect to the protein is described by the *state variables* - which are a set of variables used to describe the translation, orientation, and conformation of the ligand with respect to a protein - where each state variable corresponds to a gene. The ligand's state corresponds to the genotype and the atomic coordinates of the ligand correspond to the phenotype. The total interaction of the ligand with the protein, as determined by the energy function, defines the *fitness* of a solution. *Crossover* is the process by which random pairs of individuals (solutions) are mated, inheriting genes from either parent to produce offspring. Changes to the genes of the offspring can be introduced by random *mutation*. The current *generation's* offspring undergo *selection* based on the individual's fitness; this ensures that better solutions reproduce and that poorer solutions are terminated [220].

The Lamarckian Genetic Algorithm of AutoDock

The LGA is a hybrid search technique that combines an adaptive global optimizer, a GA, with a pseudo-Solis and Wets (pWS) LS method. In this GA the chromosome is comprised of a string of real-valued genes each of which encode one state variable. The genes are; three Cartesian coordinates which define the ligand translation, four variables that form a

quaterion which specifies the ligand orientation, and one real-value for each ligand torsion. AutoTors, an AutoDock routine, creates a torsion tree which defines the order of genes that encode the torsion angles. The ligand's state variables are therefore, one-to-one mapped to the genes of an individual's chromosome. Using real encodings to represent the genome limits the search to reasonable domains. This contrasts with the use of binary operators to represent the genome which can lead to an inefficient search by producing values outside the domain of interest [220].

Initially the LGA creates a random population of individuals, the number of which is user defined. For each individual random values are assigned to each of the genes in the following fashion: a uniformly distributed random value between the minimum and maximum *x*, *y* and *z* extents of the grid maps is assigned to each of the three *x*, *y* and *z* translation genes; a random quaterion, consisting of a random unit vector and random rotational angle between $-180\,^{\circ}$ and $+180\,^{\circ}$, is assigned to the four genes describing the orientation; and random values between $-180\,^{\circ}$ and $+180\,^{\circ}$ are assigned to the torsion angle genes. Creation of the initial population is followed by iterations of the algorithm over generations until the termination criteria have been met. Each generation consists of five processes carried out in the following order; mapping and evaluation, selection, crossover, mutation, and elitist selection. Following each generation the LS is performed on a user defined proportion of the population [220].

Mapping is performed across the entire population and translates an individual's genotype to its phenotype. Following mapping, the sum of intermolecular interaction energy between the ligand and protein, and the intramolecular interaction energy of the ligand (fitness) is calculated. The total number of energy evaluations is incremented every time an individual's energy is calculated. Proportional selection determines which individuals reproduce and ensures that individuals with better-than-average fitness receive more offspring. Determination of the number of offspring attributed to an individual is carried out in accordance with:

$$n_0 = \frac{f_w - f_i}{f_w - \langle f \rangle} \quad f_w \neq \langle f \rangle \tag{2.2}$$

where: n_0 is the integer number of offspring allocated to an individual; f_w is the fitness of the worst individual; f_i is the fitness of the individual; and $\langle f \rangle$ is the mean fitness of the population. As the numerator of this equation will always be greater than the denominator individuals of sufficient fitness will always be assigned at least one offspring. When f_w equals the mean fitness of the population the docking simulated is assumed to have converged and is terminated [220].

The number of random members of the population selected to undergo crossover and mutation is user defined. Two-point crossover is performed first and breaks are not permitted within a gene, only between genes. The offspring produced by crossover replace the parents in the population ensuring the population size remains constant. Mutation follows crossover and is performed by adding a random real number that has a Cauchy distribution to the variable. The distribution is defined by:

$$C(\alpha, \beta, x) = \frac{\beta}{\pi\left(\beta^2 + (x-\alpha)^2\right)} \qquad \alpha \geq 0, \beta > 0, -\infty < x < \infty \tag{2.3}$$

where α and β are parameters that affect the mean and spread of the distribution. Optionally an elitism parameter can be assigned to allow a user to define the number of the top individuals to be carried over to the next generation. Once one of the termination criteria is met AutoDock reports the fitness, state variables, and coordinates of the docked conformation, and carries out a conformational analysis of the docked conformations to determine which are similar. These clusters are reported ranked by increasing energy [220].

2.4 Statistical Mechanics

Statistical mechanics is the theoretical framework that allows the study of the properties of a macroscopic system and relates these to the systems' microscopic constituents. Statistical thermodynamics, a branch of the subject, is used to calculate the thermodynamic functions of a system when the interactions between the systems atoms and molecules are known or given.

Since the system of interest commonly comprises a large number of molecules and therefore a large number of mechanical degrees of freedom, the full details of their underlying dynamics are not obtainable. It is the macroscopic properties, including the thermodynamic functions, which are calculated and interpreted. These properties are gross averages over the detailed dynamic states [223]. The fundamental postulate of statistical mechanics states that "for an isolated system in equilibrium, all accessible states are equally probable" [224]. Consider a classical system containing *N* atoms. To define the state of the system each atom requires *6N* values, 3 coordinates, *r*, and 3 components of the momentum, *q*. Γ is used to denote a phase point in the 6_N-dimensional phase space, which represents each combination of *3N* positions and *3N* momenta (or quantum numbers). A system can therefore be considered as a collection of points in phase space, equation 2.4 [203].

$$\mathbf{\Gamma} = (\mathrm{r}_1, \mathrm{r}_2, \ldots, \mathrm{q}_{\mathrm{n}-1}, \mathrm{q}_{\mathrm{n}}) \tag{2.4}$$

A series of points connected in time can be generated by applying Newton's equations of motion in a simple classical system, to produce a trajectory in phase space. By following the trajectory for a long enough time and with an appropriate time step connecting the points, the system would visit all possible microstates [203,225].

If *A* is used to represent some property, e.g. the potential energy, the instantaneous value of this property as a function of Γ can be written as $A(\Gamma)$. As the system evolves, Γ and thus $A(\Gamma)$ will change. The experimentally observable 'macroscopic' property A_{obs} can be

reasonably assumed to be the time average of $A(\Gamma)$ over a long time interval, equation 2.5.

$$A_{\mathrm{obs}} = \langle A \rangle_{\mathrm{time}} = \langle A\left(\Gamma\left(t\right)\right)\rangle_{\mathrm{time}} = \lim_{\mathrm{t_{obs}}\to\infty} \frac{1}{t_{obs}} \int_0^{\mathrm{t_{obs}}} A\left(\Gamma\left(t\right)\right) \delta t \tag{2.5}$$

Whilst solving Newton's equations of motions to a desired accuracy for a system containing 1000 atoms is a practical proposition, this is not true for a macroscopic number of atoms, 10^{23}. The integration of equation 2.5 cannot be extended to infinite time, but it can be averaged over a long finite time, t_{obs}. This is done in Molecular Dynamics (MD) where the equations of motion are solved on a step-by-step basis. Equation 2.5 can be rewritten as:

$$A_{\mathrm{obs}} = \langle A \rangle_{\mathrm{time}} = \frac{1}{\tau_{\mathrm{obs}}} \sum_{\tau=1}^{\tau_{obs}} A\left(\Gamma\left(\tau\right)\right) \tag{2.6}$$

where τ_{obs} defines the number of time steps and $\delta\ t = t_{obs}\tau_{obs}$ defines the length of the time step. This calculation of the time average is in principle straightforward, however the complexity of the time evolution of $A(\Gamma(t))$ for 'macroscopic' systems is such that it is replaced by the ensemble average. The collection of points, Γ, in phase space that comprise the ensemble and satisfy the conditions of a particular thermodynamic state, are distributed according to a probability density $\rho(\Gamma)$. The chosen fixed macroscopic parameters determine the function $\rho(\Gamma)$ and the general notation, ρ_{ens} , is used to reflect this. At any chosen instant in time, each point represents a typical configuration of the system. According to Newton's laws of motion each system evolves in time independently of all other systems. As a result, the phase space density, $\rho_{ens}\left(\Gamma\right)$, changes with time. During this evolution no systems are created or destroyed. The time-dependence completely disappears in an equilibrium ensemble, $\rho_{ens}\left(\Gamma\right)$. During the system evolution, as each system leaves one state $\mathbf{\Gamma}(\tau)$ and moves onto the next, $\mathbf{\Gamma}(\tau+1)$, another system arrives, $\mathbf{\Gamma}(\tau-1)$, to replace the last. If just one trajectory passes through all non-zero ρ_{ens} points in phase space then it follows that each system will eventually visit all state points. This system is termed 'ergodic' and

the majority of many-particle systems in nature are ergodic. For a many-bodied system the time taken to complete a cycle is immeasurably long so the time average is replaced by a snapshot of the system, which is an average over all the ensemble members at that point in time. Such a snapshot is termed the ensemble average, equation 2.7 [226].

$$A_{\mathrm{obs}} = \langle A \rangle_{\mathrm{ens}} = \langle A \left| \rho_{\mathrm{ens}} \right| \rangle = \sum_{\mathrm{r}} A\left(\Gamma\right) \rho_{\mathrm{ens}}\left(\Gamma\right) \tag{2.7}$$

When the ensemble average is calculated over a long time and many microstates, it is equivalent to the time average:

$$\langle A \rangle = A_{\mathrm{obs}} \tag{2.8}$$

Four commonly used ensembles are:

- Microcanonical - constant NVE
- Canonical - constant NVT
- Isothermal-isobaric - constant NPT
- Grand canonical - constant μVT

where the variables are the number of particles *N*, the volume *V*, the energy *E*, the temperature *T*, the pressure *P* and the chemical potential μ. In each ensemble the thermodynamic variables are fixed. In biological MD simulations the most commonly used ensemble is the isothermal-isobaric, since this corresponds to experiments where the surroundings define the temperature and pressure (*i.e.* a thermodynamically open system).

2.5 Molecular Dynamics

A wealth of information can be gained by studying the dynamical behaviour of a system as a function of time. This can be done using Molecular dynamics, MD. Successive configurations of the system are generated by integrating Newton's equations of motion, equation 2.9. This produces a trajectory that shows how the position of the atoms and momenta evolve with time. Newton's second law of motion states that force equals the acceleration, $F = ma$. The trajectory is obtained by solving the differential equations embodied within this law.

$$\frac{d^2x_i}{dt^2} = \frac{F_{xi}}{m_i} \tag{2.9}$$

where m_i is the mass of the particle moving in the direction of coordinate x_i with force acting upon the particle in the direction of x_i, F_{xi}.

In simulations of intermolecular interactions of realistic models, the force acting on each atom is dependent on its position relative to that of the other atoms. Such a simulation uses continuous potentials. Under the influence of a continuous potential the motion of such systems is difficult to describe analytically as the coupled nature of the particles gives rise to a many-body problem. It such circumstances it is necessary to use a finite difference method to integrate the equations of motion.

2.5.1 Finite Difference Methods

Finite difference methods are used to generate MD trajectories for continuous potential models by breaking down the integration into small steps separated by a fixed time, δt. The choice of this time interval will be discussed in more detail later. From the positions and velocities of each particle in the system at time t, the positions and velocities of the particles at time $t + \delta t$ can be calculated. The force on each particle within the system is

calculated as the vector sum of the force exerted by every other particle. The accelerations of the particles are calculated from the forces and masses, according to Newton's second law. The accelerations are combined with the positions and velocities at time *t* to generate the positions and velocities at time *t* + δt. The forces on the particles can then be calculated from their new positions, the accelerations using the forces and the positions and velocities calculated at time *t* + $2\delta t$, and the process iterated up to the total desired duration of the trajectory. All finite difference algorithms for integrating the equations of motion are based on the assumption that an estimate of the positions and dynamic properties at any time *t* + δt can be approximated using the Taylor series expansion about time *t*:

$$\mathbf{r}\left(t+\delta t\right)=\mathbf{r}\left(t\right)+\delta t\mathbf{v}\left(t\right)+\frac{1}{2}dt^{2}\mathbf{a}\left(t\right)+\frac{1}{6}dt^{3}\mathbf{b}\left(t\right)+\frac{1}{24}dt^{4}\mathbf{c}\left(t\right)+\ldots \tag{2.10}$$

$$\mathbf{v}\left(t+\delta t\right)=\mathbf{v}\left(t\right)+\delta t\mathbf{a}\left(t\right)+\frac{1}{2}dt^{2}\mathbf{b}\left(t\right)+\frac{1}{6}dt^{3}\mathbf{c}\left(t\right)+\ldots \tag{2.11}$$

$$\mathbf{a}\left(t+\delta t\right)=\mathbf{a}\left(t\right)+\delta t\mathbf{b}\left(t\right)+\frac{1}{2}\delta t^{2}\mathbf{c}\left(t\right)+\ldots \tag{2.12}$$

$$\mathbf{b}\left(t+\delta t\right)=\mathbf{b}\left(t\right)+\delta t\mathbf{c}\left(t\right)+\ldots \tag{2.13}$$

where **r** represents the positions, **v** the velocity (first derivative), **a** the acceleration (second derivative), **b** is the third derivative etc.

2.5.2 The Verlet Algorithm

The Verlet algorithm is the most widely used implementation of the finite difference method. It is efficient, stable, has modest storage requirements and is easy to implement. In this algorithm the positions and accelerations at time *t* and those from the previous step, **r**(*t* - δt), are used to calculate the new positions at time *t* + δt. The relationship between the positions and accelerations, and the velocities can be written as:

$$\mathbf{r}\left(t+\delta t\right)=\mathbf{r}\left(t\right)+\delta t\mathbf{v}\left(t\right)+\frac{1}{2}\delta t^{2}\mathbf{a}\left(t\right)+\ldots \tag{2.14}$$

$$\mathbf{r}\left(t-\delta t\right)=\mathbf{r}\left(t\right)-\delta t\mathbf{v}\left(t\right)+\frac{1}{2}\delta t^{2}\mathbf{a}\left(t\right)-\ldots \tag{2.15}$$

These equations can be combined to give:

$$\mathbf{r}\left(t-\delta t\right)=2\mathbf{r}\left(t\right)-r\left(t-\delta t\right)+\delta t^{2}\mathbf{a}\left(t\right) \tag{2.16}$$

Disadvantages of the Verlet algorithm include the difficulty with which velocities can be obtained due to the lack of an explicit velocity term. Indeed velocities can only be obtained once the positions have been determined for the next step. Whilst velocities are not needed to compute the trajectory, they are a useful estimate of kinetic energy. They can be obtained using the following equation:

$$\mathbf{v}\left(t\right)=\frac{\left[\mathbf{r}\left(t+\delta t\right)-\mathbf{r}\left(t-\delta t\right)\right]}{2\delta t} \tag{2.17}$$

Velocities calculated using this equation are subject to errors in the order of δt^2.

Another disadvantage is the loss of precision produced when obtaining the positions **r**(t + δt). Here it is necessary to add a small term, (δt^2**a**(t)), to the difference between two larger terms, **2r**(t) and **r**(t - δt) [203,226].

2.5.3 The Leap-frog Algorithm

Several variations of the Verlet algorithm exist and include the velocity Verlet and leap-frog algorithms [227]. GROMACS, which was used to produce the simulations presented within this thesis, uses the leap-frog algorithm to integrate the equations of motion. The following relationships are used in the leap-frog algorithm:

$$\mathbf{r}\left(t+\delta t\right)=\mathbf{r}\left(t\right)+\delta t\mathbf{v}\left(t+\frac{1}{2}\delta t\right) \tag{2.18}$$

$$\mathbf{v}\left(t+\frac{1}{2}\delta t\right)=\mathbf{v}\left(t\right)-\frac{1}{2}\delta t+\delta t\mathbf{a}\left(t\right) \tag{2.19}$$

Firstly the velocities **v**(t + 1/2t) are calculated from the velocities at time t - 1/2δt and the accelerations at time t, equation 2.19. The newly calculated velocities and positions at time **r**(t) are then used to determine the positions at time t + δt, equation 2.18. At time t the velocities can be calculated using:

$$\mathbf{v}\left(t\right)=\frac{1}{2}\left[\mathbf{v}\left(t+\frac{1}{2}\delta t\right)+\mathbf{v}\left(t-\frac{1}{2}\delta t\right)\right] \tag{2.20}$$

The velocities leap over the coordinates to give the next half-step values at time t + 1/2δt. In turn the positions leap over the velocities to give their new coordinates at time t + δt.

At each stage the current positions $\mathbf{r}(t)$ and accelerations $\mathbf{a}(t)$ are stored together with the half-step velocities $\mathbf{v}(t - 1/2\ \delta t)$). The leap-frog algorithm, although compensating some of the disadvantages of the Verlet algorithm, has the disadvantage that the positions and velocities are not synchronised. As a result it is not possible to calculate the contribution of the kinetic energy to the total energy at the same time that the positions are defined.

Important considerations when choosing the integration algorithm is that it is time-reversible, conserves energy and momentum and will allow the use of a long time step δt [228].

2.5.4 Time Step

The choice of time step must balance the need to generate a 'correct' trajectory and one that covers a sufficient proportion of phase space. If the time step is too small the trajectory will not sample enough of the phase space. A time step that is too large may produce instabilities in the integration algorithm as a direct result of atoms overlapping and causing regions of high energy. Violation of energy would be caused as a result of these instabilities and thus numerical overflow would cause the program to fail.

To maintain numerical stability, a general rule is that the time step should be one to two orders of magnitude smaller than the fastest periodic motion within the system. The fastest motion within a classical system is bond vibration. For a heavy-atom-hydrogen bond this is 10^{-14} s, therefore the time step for a simulation containing such bonds should generally not exceed 0.1 femtoseconds (fs). This short time step clearly imposes limitations on the attainable length of large-scale MD simulations. The time step can be increased when such vibrations are frozen out by constraining bond lengths to their optimal values. This is a reasonable assumption as the amplitude of C-H bond vibrations is small with respect to other atom-atom distances at normal temperatures and such bond vibrations would have little effect on the behaviour of the overall system. Bond lengths can be fixed using an appropriate constraint algorithm, *i.e.* SHAKE [229] or LINCS [230].

2.6 Setting up a Molecular Dynamics Simulation

A set of coordinates is required as input into the simulation. The system is then evolved from the starting coordinates during an equilibration phase where structural and thermodynamic properties are closely monitored until a plateau is reached. After equilibration, a production run is performed during which simple properties of the system are calculated. The configuration of the system is saved at regular intervals. Finally, post simulation analysis is performed and the output configurations studied. Unusual changes in the structure of the system can highlight abnormalities in the simulation.

2.6.1 The Initial Configuration

Initial coordinates of the system can be taken from experimental data, *e.g.* NMR or X-Ray crystallography, generated by theoretical modelling or a combination of both. The choice of this configuration is critical as this it can determine the success of the simulation. High-energy interactions, which may cause instabilities in the simulation, can be eradicated by performing an energy minimisation before the simulation.

2.6.2 Energy Minimisation

Energy minimisation algorithms are used to identify the geometries of a system that correspond to minimum points on the energy surface. Two first-order, first derivative, minimisation algorithms which are commonly used in molecular modelling are Steepest Descent (SD) and Conjugate Gradient (CG). Both methods gradually change the coordinates of the system as they move down a gradient bringing the structure closer to the minimum. The starting set of coordinates for each iteration is the molecular configuration generated by the previous step.

Steepest Descents moves in the direction parallel to the net force. For *3N* coordinates the direction of movement is represented by the *3N*-dimensional unit vector r. The maximum

displacement h_0 defines how far to move along the gradient. The forces F and potential energy are calculated and new positions are computed using equation 2.21.

$$rn + 1 = r_n + \frac{F_n}{\max(|F_n|)} h_n \tag{2.21}$$

where $\boldsymbol{F}_n$ is the force or negative gradient of the potential V and h_n the maximum displacement. The largest absolute values of the force components are denoted *max* ($|\boldsymbol{F}_n|$). The forces and potential energy of the new positions are then determined and the new positions accepted or rejected based on a set of criteria. In GROMACS positions are accepted when ($V_{n+1} < V_n$) and h_n+1 then becomes 1.2 h_n. Positions are rejected when (V_n+1 $> V_n$) and h_n becomes 0.2 h_n. The search ceases when either the number of user specified force evaluations have been performed or when *max*($|\boldsymbol{F}_n|$) is less than a specified value 0.

Steepest Descents is a good method for relieving the highest-energy interactions in an initial structure, as the direction of the gradient is determined by the largest interatomic forces. Even when the starting configuration is far from the energy minimum, where the harmonic approximation of the energy surface is often a poor assumption, this method is very robust. It is also easy to implement. However, if the minimum is located at the base of a long narrow valley a vast number of very small steps will be required to obtain convergence. At each step a right-angles turn is required, generating an oscillating path which continually overcorrects itself. During the later stages of the minimisation errors corrected by earlier moves are reintroduced [228].

The conjugate gradient method is slower than the steepest descent in the initial minimization stages, but is more efficient the closer the structure is to the energy minimum. The path generated by the conjugate gradient algorithm in narrow valleys does not exhibit the oscillatory behaviour of the steepest descents method, as although the gradients of each step are orthogonol the directions are conjugate 'M' steps. This method moves from point x_k in direction v_k where v_k is calculated from the gradient at the point and the previous direction, v_k -1.

$$v_k = -\mathbf{g}_k = \gamma_k v_{k-1} \tag{2.22}$$

where γ_k is a scalar constant [231].

A short run of steepest descent minimsation followed by conjugate gradients minimisation can be used to achieve a relaxed starting structure.

2.6.3 Generating the Initial Velocities

Next, initial atomic velocities, if not available, must be assigned. This is usually done by randomly selecting from a either a Maxwell-Boltzmann or Gaussian distribution at the required temperature, corrected so that there is no overall momentum. The Maxwell-Boltzmann equation, equation 2.23, can be used to obtain the probability density for the velocity component v_{ix} at given absolute temperature T for an atom I of mass in the direction x.

$$p\left(v_i\right) = \sqrt{\frac{m_i}{2\pi k_B T}} \exp\left(-\frac{m_i v_i^2}{2k_B T}\right) \tag{2.23}$$

where k_B is Boltzmann's constant. Similar equations apply in the y and z directions. A random seed is used to generate the first set of velocities which for the leap-frog algorithm are at $t = t_0 - \Delta t/2$. In GROMACS normally distributed random numbers are generated by adding twelve random numbers R_k in the range $0 = R_k = 1$, and subtracting 6.0 from their sum, this is then multiplied by the standard deviation of the velocity distribution. A correction is then made to ensure the resulting total energy will correspond to the desired temperature. The center-of-mass motion is removed and velocities are scaled to ensure the required temperature, T, is obtained.

2.6.4 Equilibration

The starting point for a simulation is often at a different density or temperature to that required. Therefore, it is necessary to run the simulation for a period to allow the system to come to equilibrium. At the end of this phase of the simulation all memory of the starting configuration should have been lost. Various parameters - including the temperature, energy and pressure of the system in addition to the configuration of system - can be monitored during this phase. When these parameters cease to exhibit a systematic drift and start to oscillate about steady mean values the production run can commence.

2.7 Periodic Boundary Conditions

When simulating bulk liquids the correct treatment of boundaries and boundary effects is crucial, as it enables a simulation on a relatively small number of particles to be used to calculate properties on a macroscopic level. Applying periodic boundary conditions (PBC) is the classical way to minimise edge effects in a finite system [232]. The atoms of the system to be simulated are placed into a box, which is replicated in three dimensions to form an infinite lattice, see figure 2.2. When simulating a crystal, such boundary conditions are required. When simulating a non-periodic system, such as a liquid, this periodicity can cause errors; however these errors are thought to be less severe than the errors resulting from an unnatural boundary with a vacuum, and can be evaluated by comparing various system sizes.

In principle, the unit cell can be one of several shapes. Commonly used shapes include cubic, rectangular, rhombic dodecahedron and the truncated octahedron. The later two are closer in shape to a sphere and are therefore more computationally economical for an isotropic liquid. The volume of a rhombic dodecahedron is 71% that of the equivalent cube saving approximately 29% of CPU-time when simulating a macromolecule in solvent.

Periodic boundary conditions are commonly employed in combination with use of the minimum image convention. Using this method, only the closest image of the atom is considered

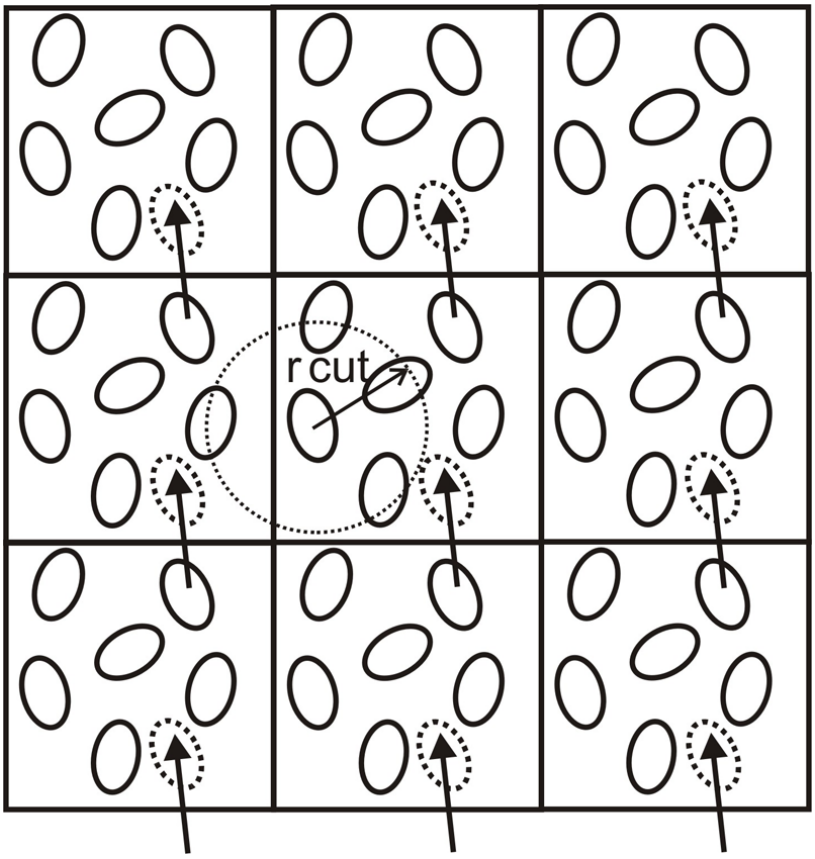


Figure 2.2: **A two dimensional schematic of periodic conditions.** Image adapted from [226].

when evaluating short-range non-bonded interactions within a user defined cut-off. When a cut-off is used, interactions between all pairs of atoms outside this region are set to zero. The chosen cut-off should not be so large as to allow a particle to 'see' its own image or the same molecule twice. For systems in which the angels do not deviate significantly from orthorombic, this can be expressed as:

$$R_c < \frac{1}{2}\min\left(\|\mathbf{a}\|, \|\mathbf{b}\|, \|\mathbf{c}\|\right), \tag{2.24}$$

where the cutoff radius must not exceed half the shortest box vector. For solvated macromolecule simulations a further restriction is applied by the desire to not allow a single solvent molecule 'see' both sides of the macromolecule. Therefore, when defining the size of the box, a general rule to follow is that the box needs to be at least the length of the macromolecule plus twice the length of the short range cut-off R_c [228].

2.8 Force Fields

In MD the force acting on an atom can be written as:

$$F_i = -\nabla U \tag{2.25}$$

A force field comprises a set of equations and a set of parameters to be used in these equations. The set of equations, or potential functions, is used to generate the potential energy, U, which gives rise to the forces. Potential functions can be divided into two main categories: Bonded and Non-bonded. Bonded interactions include covalent bond stretching, angle-bending, improper and proper dihedrals. Non-bonded interactions include van der Waals short range contributions and an electrostatic contribution. When using a set of equations and a set of parameters, care must be taken to assure the combination forms a consistent set. Parameter sets are often developed over long time periods to reproduce a set of experimental values. For this reason and as contributions to the total force are usually interdependent, any changes to the individual parameters of a given force field must be made with caution.

The simulations presented in this thesis where performed using the GROMOS96 force field 43a2 parameter set [233] implemented in the Groningen Machine for Chemical Simulations (GROMACS) [234,235] suite of molecular simulation programs. Gromacs is a united atom force field. Only polar hydrogen atoms and those in aromatic systems are explicitly modelled. Other empirical force fields widely used for biomolecular simulations include CHARMM, AMBER and OPLS. For a review of these empirical force fields see [236].

A class one additive potential energy function is commonly used in biomolecular force fields to describe the relationship between energy and structure:

$$U_{total} = (U_{bonds} + U_{angles} + U_{dihedral} + U_{improper}) + (U_{nonbonded}) \tag{2.26}$$

2.8.1 Bonded Interactions

Bonded interactions include: bond stretching, a 2-body interaction; bond angles, a 3-body interaction; and dihedral and improper dihedral angles, a 4-body interaction. In GROMACS all bonded interactions are calculated based on a fixed atom list.

2.8.2 Bond Stretching Potential

The bond stretching potential is used to describe explicit covalent bonds between specified pairs of atoms. Whilst a Morse potential more accurately reproduces the anharmonic nature of a covalent bond it is not commonly used, as it is rare for bond lengths to deviate significantly from their equilibrium values in MD simulations of biological molecules. Instead a harmonic potential is used:

$$U_{bond}\left(r_{ij}\right) = \frac{1}{2}{k^b}_{ij}\left(r_{ij} - {r_{ij}}^0\right)^2 \tag{2.27}$$

where k^b is the force constant associated with bond r and r^0 is the reference or ideal bond length.

2.8.3 Angle Potential

A harmonic potential can also be used to represent the bond angle vibration between a triplet of atoms, *i*-*j*-*k*:

$$U_{angle}\left(\theta_{ijk}\right) = \frac{1}{2} k_{ijk}^{\theta} \left(\theta_{ijk} - \theta_{ijk}^{0}\right)^{2} \tag{2.28}$$

where k^{θ} is the force constant associated with angle θ_{ijk} and atom *j* is the middle atom in a sequence of covalently bonded atoms.

The GROMOS-96 force field contains a computationally more efficient function to represent angle vibrations:

$$V_{\alpha}\left(\theta_{ijk}\right) = \frac{1}{2} k_{ijk}^{\theta} \left(\cos\left(\theta_{ijk}\right) - \cos\left(\theta_{ijk}^{0}\right)\right)^{2} \tag{2.29}$$

where:

$$\cos\left(\theta_{ijk}\right) = \frac{\mathbf{r}_{ij} - \mathbf{r}_{kj}}{r_{ij} r_{kj}} \tag{2.30}$$

Partial differentiation with respect to the atomic positions yields the corresponding force. In this function, the force constants are related to the force constants in the harmonic angle potential, $k^{\theta,harm}$, by:

$$k^{\theta} \sin^{2}\left({\theta_{ijk}}^{0}\right) = k^{\theta,harm} \tag{2.31}$$

2.8.4 Proper Dihedrals

In GROMACS there is a choice of either the periodic function or a function based on the powers of *cos*φ, the Ryckaert-Bellemans potential, to describe the normal dihedral interactions.

2.8.5 Proper Dihedrals: Periodic Function

The IUPAC/IUB convention is applied when defining the proper dihedral angle. The angle φ is that between the *i-j-k* and the *j-k-l* planes.

$$U_{proper}\left(\varphi_{ijkl}\right) = k\varphi\left(1 + \cos\left(n\varphi - \varphi_0\right)\right) \tag{2.32}$$

When using this potential a special 1-4 Lennard Jones interaction must be included.

2.8.6 Proper Dihedrals: Ryckaert-Bellemans Function

This function is often used for alkanes:

$$U^{RB}_{proper}\left(\varphi_{ijkl}\right) = \sum_{n=0}^{5} C_n \left(\cos\left(\psi\right)\right)^n \tag{2.33}$$

where ψ = φ - 180$\circ$.

Lennard-Jones interactions between the first and last atom of the dihedral must be excluded when using this potential and ψ is defined according to the 'polymer convention' (ψ_{trans} = 0).

2.8.7 Improper Dihedrals

Improper dihedral angle potentials are used to either ensure planar groups remain in the plane or to prevent structures from flipping into their mirror images.

$$U_{improper}\left(\zeta_{ijkl}\right) = k_{\zeta}\left(\zeta_{ijkl} - \zeta_0\right)^2 \tag{2.34}$$

This is a harmonic potential and periodicity is not taken into account, therefore it is best to define ζ_0 as far away from $\pm$ 180° as possible.

2.8.8 Non-Bonded Interactions

Using a neighbour-list can significantly reduce the time taken to compute non-bonded interactions. This is a reasonable assumption as an atom's neighbours, those within the cut-off distance, do not change significantly over 10 to 20 time steps. Throughout the simulation the neighbour-list is updated at regular intervals. The distance used to calculate the list should be equal to or larger than the actual non-bonded cut-off distance. This is so no atom initially outside the neighbour-list cut-off, approaches closer than the non-bonded cut-off distance before the neighbour list can be updated [231].

In GROMACS the non-bonded interactions are pair-additive, centro-symmetric and calculated using a neighbour list.

$$U\left(\boldsymbol{r}_1, \ldots \boldsymbol{r}_N\right) = \sum_{i<j} V_{ij}\left(\boldsymbol{r}_{ij}\right); \tag{2.35}$$

$$\boldsymbol{F}_i = -\sum_j \frac{dV_{ij}\left(r_{ij}\right)}{dr_{ij}} \frac{\boldsymbol{r}_{ij}}{r_{ij}} = -\boldsymbol{F}_j \tag{2.36}$$

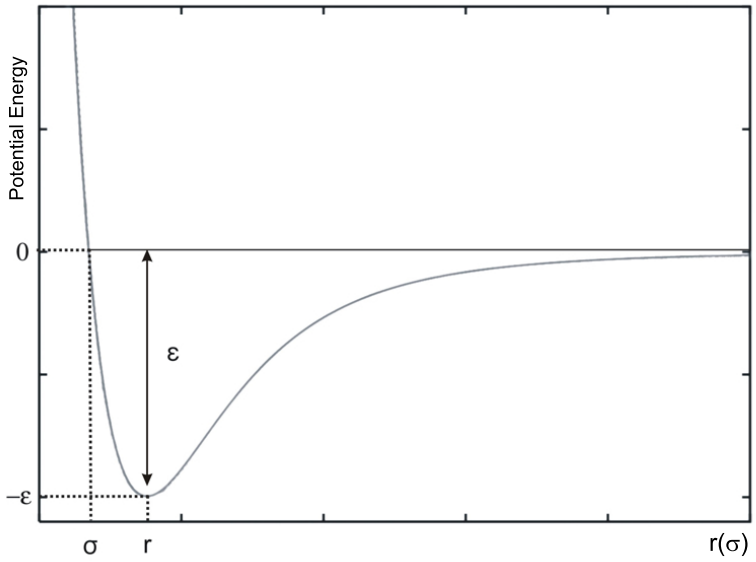


Figure 2.3: **Schematic representation of the Lennard-Jones potential function.** Image adapted from [237].

The non-bonded interactions comprise repulsion and dispersion terms combined in the Lennard-Jones (6-12 interaction) potential and a Coulomb term through which partially charged atoms act.

2.8.9 Lennard-Jones nteraction

The Lennard-Jones expression, U_{LJ}, is used to describe the short ranges interactions produced by van der Waals forces (vdW):

$$U_{LJ}\left(r_{ij}\right) = 4\epsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right) \tag{2.37}$$

where the bond length, σ, and the well depth, ϵ, depend on pairs of atom types and are commonly taken from a matrix of LJ-parameters.

This potential is represented in schematic 2.3:

2.8.10 Coulomb Interaction

Equation 2.38 describes the Coulomb interaction between two charged particles:

$$U_c\left(r_{ij}\right) = f\frac{q_i q_j}{\epsilon_r r_{ij}} \tag{2.38}$$

where:

$$f = \frac{1}{4\pi\epsilon_0} \tag{2.39}$$

From this potential the force derived is:

$$\boldsymbol{F}_i\left(\boldsymbol{r}_{ij}\right) = f\frac{q_i q_j}{\epsilon_r r_{ij}^2}\frac{\boldsymbol{r}_{ij}}{r_{ij}} \tag{2.40}$$

2.8.11 Coulomb Interaction: Particle-Mesh Ewald

The long-ranged electrostatic forces were calculated using the particle-mesh Ewald (PME) method [238] [239]. This method, proposed by Tom Darden, is designed to improve the performance of the reciprocal sum over that in the conventional Ewald method [240].

The Ewald method converts the single slowly-converging sum, equation 2.41, into quickly-converging terms and a constant term. The total charge-charge contribution to the electrostatic energy from *N* particles and their periodic images is given by:

$$V = \frac{f}{2} \sum_{n_x} \sum_{n_y} \sum_{n_{z^*}} \sum_{i}^{N} \sum_{j}^{N} \frac{q_i q_j}{\mathrm{r}_{ij,\mathrm{n}}} \tag{2.41}$$

where (n_x, n_y, n_z) = n, is the box index vector. The star describes that terms where $i = j$ should not be included when the box index vector = 0, 0, 0. r_{ij}, n indicates the minimum distance between the charges i and j.

$$V = V_{dir} + V_{rec} + V_0 \tag{2.42}$$

$$V_{dir} = \frac{f}{2} \sum_{ij}^{N} \sum_{n_x} \sum_{n_y} \sum_{n_z^*} q_i q_j \frac{\mathrm{erfc}\left(\beta r_{ij,\mathrm{n}}\right)}{r_{ij,\mathrm{n}}} \tag{2.43}$$

$$V_{rec} = \frac{f}{2\pi V} \sum_{ij}^{N} q_i q_j \sum_{m_x} \sum_{m_y} \sum_{m_{z^*}} \frac{\exp\left(-\left(\pi \mathrm{m}/\beta\right)^2 + 2\pi i \mathrm{m} \cdot \left(\mathrm{r}_i - \mathrm{r}_j\right)\right)}{\mathrm{m}^2} \tag{2.44}$$

$$V_0 = \frac{-f\beta}{\sqrt{\pi}} \sum_{i}^{N} q_i^2, \tag{2.45}$$

The relative weight between the direct space sum and the reciprocal space sum is determined by parameter β and m = (m_x, m_y, m_z). By doing this, the direct space sum will use a short cutoff (1 nanometre, nm) as will the reciprocal space sum. It is not viable to use the Ewald method in larger systems as the computational cost of the reciprocal part of the sum increases as N^2 or sometimes $N^{3/2}$.

In PME the wave vectors are not directly summed. Using Cardinal B-spline interpolation the wave vectors are instead assigned to a grid, which is then Fourier transformed, using

fast-Fourier-transform methods, and the reciprocal energy term calculated by a single sum over the grid in k-space. This algorithm scales as *N* log(*N*) and as a result is substantially faster than the conventional Ewald method when simulating larger systems.

Calculating the long-range interactions with PME dictates that the short-range coulomb potential be modified. The short range potential then becomes:

$$U(r) = f\frac{\operatorname{erfc}(\beta r_{ij})}{r_{ij}} q_i q_j, \tag{2.46}$$

The relative weight between the direct space sum and the reciprocal space sum is determined by parameter β. The complementary error function is erfc(*x*).

2.8.12 Special Interactions: Position Restraints

Position restraints are used to tether particles to their reference position R_i. Position restraints can be used to restrain the motion of a molecule while solvent is equilibrated or too restrain a shell of particles around a region that is simulated in detail. The restraint potential form is:

$$U_{pr}(r_i) = \frac{1}{2} k_{pr} \left| r_i - \boldsymbol{R}_i \right|^2 \tag{2.47}$$

This can be rewritten and the forces expressed as:

$$V_{pr}(r_i) = \frac{1}{2}\left[k_{pr}^x (x_i - X_i)^2 \hat{\mathrm{x}} + k_{pr}^y (y_i - Y_i)^2 \hat{\mathrm{y}} + k_{pr}^z (z_i - Z_i)^2 \hat{\mathrm{z}}\right] \tag{2.48}$$

$$F_i^x = -k_{pr}^x \left(x_i - X_i\right) \tag{2.49}$$

$$F_i^y = -k_{pr}^y \left(y_i - Y_i\right) \tag{2.50}$$

$$F_i^z = -k_{pr}^z \left(z_i - Z_i\right) \tag{2.51}$$

The use of three different force constants allows the position restraints to be turned on or off in each dimension, facilitating the harmonic restraint of atoms to one plane if required.

2.8.13 LINCS

The LINCS algorithm within GROMACS resets bonds to their correct lengths [230]. The method uses two steps and therefore is non-iterative. No matrix-matrix multiplications are needed even though LINCS is based on matrices. The advantages of using LINCS rather than the established SHAKE algorithm include that it is faster and more stable, however it can only be used with bond constraints and isolated angle constraints.

In a system of *N* particles where the positions are given by a *3N* vector, *r*(*t*), the equations of motion according to Newtons's law are:

$$\frac{\mathrm{d}^2 r}{\mathrm{d}t^2} = M^{-1} F \tag{2.52}$$

where *M* is a *3N* x *3N* diagonal matrix containing the particle masses and *F* the *3N* force vector. *K* time-independent constraint equations serve to constrain the system:

$$g_i(r) = |r_{i_1} - r_{i_2}| - d_i = 0 \quad i = 1, \ldots, K \tag{2.53}$$

2.8.14 Simple Point Charge Water

The Simple Point Charge, SPC, water model was developed in 1982. It is a tetrahedral, three site model of water. The O-H distance is fixed at 0.1 nanometers (nm) and the H-O-H bond angle is 109.42○. There are point charges on the hydrogen positions of +0.41e and on the oxygen -0.82 . The Lennard-Jones interaction on the oxygen atom is given by:

$$U_{LJ} = \left(\frac{A}{r}\right)^6 + \left(\frac{B}{r}\right)^{12} \tag{2.54}$$

Where A = 0.37122 (kJ/mol)$^{1/6}$ · nm and B = 0.3428 (kJ/mol)$^{1/12}$· nm [241].

2.9 Temperature Coupling

The two main methods for controlling the temperature of a system during a simulation are the weak coupling scheme of Berendsen [242], and the extended ensemble Nos$\acute{e}$-Hoover scheme [243,244]. The Berendsen scheme has the distinct advantage of being very efficient at relaxing a system to the target temperature. Once at the desired temperature it may be more important to probe a correct canonical ensemble. Whilst this is not the case when using the Berendsen algorithm the difference is usually minor.

As the Berendsen scheme has been used to perform the simulations presented in this thesis, it will be considered in further detail. The Berendsen algorithm mimics the effect of coupling the system to an external heat bath with temperature T_0 via first order kinetics. A correction is applied to the system to accommodate any deviation of the system from the

desired temperature, as a result the temperature deviation decays exponentially with a time constant *t*:

$$\frac{\mathrm{d}T}{\mathrm{d}t} = \frac{T_0 - T}{\tau} \tag{2.55}$$

The Berendsen algorithm allows the user to adapt the strength of the coupling as required. The coupling time can be short for equilibriation purposes (e.g. 0.01 ps) or for more robust reliable equilibrium runs the coupling constant can be lengthened to, for example, 0.5 ps. Using a longer time coupling constant hardly influences the conservative dynamics at all.

At each step the velocities of each particle are scaled with a time-dependent factor λ. This is how the flow of heat into and out of the system is controlled. λ is defined as:

$$\lambda = \left[1 + \frac{\Delta T}{\tau T}\left\{\frac{T_0}{T\left(t - \frac{\Delta T}{2}\right)} - 1\right\}\right]^{1/2} \tag{2.56}$$

where τT, the temperature coupling time constant, is related to the time constant τ of the temperature coupling by:

$$\tau = \frac{2C_V\tau_T}{N_{df}k_B} \tag{2.57}$$

The total heat capacity of the system is represented by C_V, Boltzmann's constant k_B and the total number of degrees of freedom N_{df}. The change in temperature is less than the scaling energy, $\tau \neq \tau T$, as the kinetic energy change caused by rescaling the velocities is partly distributed between the kinetic and potential energy. In normal use λ will always be close to 1.0; however experience suggests values in the range $0.8 \leq \lambda \leq 1.25$ are reasonable, but

that values outside this range could cause a simulation to crash [228].

If a system comprising protein and water is coupled as a single unit to one heat bath, it is likely that the water will heat up and the protein cool down, leading to a temperature difference between the two groups of up to 100K. This is as a result of inefficient energy exchange between the different components, mainly due to effects such as cutoffs. The use of a proper electrostatic method can reduce these differences but not by enough to make their effect negligible. This problem can be avoided by coupling the groups separately to heat baths. For a system containing protein, ligand, water and counter ions three heat baths would be specified as the water and counter ions would be defined as one group and coupled to one heat bath [228].

2.10 Pressure Coupling

Methods for coupling a system to a 'pressure bath' include the Berendsen algorithm [242] and the extended ensemble Parrinello-Rahman approach [245]. The choice of temperature coupling method does not affect which pressure coupling method can be used. The Berendsen coupling method has been used to control the pressure in these simulations and therefore shall be discussed further.

Using the Berendson algorithm the coordinates and box vectors are scaled at every time step with a matrix μ. This has the effect of relaxing the pressure towards the reference value $\mathbf{P}_0$ as per first-order kinetics:

$$\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}t} = \frac{\mathbf{P}_0 - \mathbf{P}}{\tau_p} \tag{2.58}$$

where μ is derived from:

$$\mu_{ij} = \delta_{ij} - \frac{\Delta t}{3\tau_p}\beta_{ij}\left\{P_{0ij} - P_{ij}\left(t\right)\right\} \tag{2.59}$$

and β represents the isothermal compressibility of the system. This will most likely be a diagonal matrix, the diagonal elements of which are equal, which usually has an unknown value. A rough estimate can be used as the average pressure of the system is not affected, only the non-critical time constant of the pressure relaxation is influenced by the value of β. At 1 atmospheres (atm) and 300 Kelvin (K), β = 4.6 x 10^{-5} Bar^{-1} for water and other liquids have comparative values. In order to fulfill the box restriction, the system must be rotated when scaling anisotropically. The actual scaling matrix μ' is:

$$\mu' = \begin{pmatrix} \mu_{xx} & \mu_{xy} + \mu_{yx} & \mu_{xz} + \mu_{zx} \\ 0 & \mu_{yy} & \mu_{yz} + \mu_{zy} \\ 0 & 0 & \mu_{zz} \end{pmatrix} \tag{2.60}$$

where scaling or rotation is applied to the velocities. Slower scaling or a smaller timestep may be required when using full anisotropic deformations to prevent errors arising in the constraint algorithms.

The system can also be scaled isotropically using Berendsen coupling. In this case *P* is replaced by a diagonal matrix with elements of size trace(*P*)/3. Semi-isotropic scaling can be used when simulating interfaces. Here the *x*/*y*-directions are scaled isotropically and the *z* direction scaled independently. In either direction the scaling can be set to zero and scaling only applied in the other direction [228].

2.11 Molecular Dynamics of Biomolecules

The first biomolecular MD simulation, of bovine pancreatic trypsin inhibitor (BPTI), was published in 1977 [246]. The simulation was performed in vacuum with a simple molecular mechanics force field and the length of the simulation was 9.2 ps. This simulation, on a relatively small protein of 58 residues, transformed the view of proteins from that of relatively rigid structures to fully dynamic molecules with motions that play a functional role [247].

MD simulation is now a standard tool for the study of biomolecules, complementary to experimental techniques. The application of biomolecular simulation can be divided into three main areas.

Firstly, MD simulations of biological molecules can be used to provide details of the natural dynamics of a system in solution over a range of timescales. Specific questions about the properties of a system can be accessed using MD simulations of a model system often more easily than through actual experiments. Observations obtained from the simulation can then by validated using experimental techniques. The accuracy of such simulations can be assessed in this way and criteria obtained to improve biomolecular simulation methodology.

Secondly, thermal averages of molecular properties can be obtained from MD simulations. Using the ergodic hypothesis, the bulk properties of fluids and the free energy differences of processes such as ligand binding can be calculated.

Thirdly, the thermally accessible conformations of a molecule or complex can be explored. In ligand-docking applications, this technique can be used to explore conformational space. Another example of this application includes the use of data obtained from experiments, in the form of restraining potentials, in combination with MD and often simulated annealing protocols, to determine or refine structures [247,248].

Currently, simulations are routinely performed on systems of 10^4 atoms on nanosecond timescales. Much of the increase in computing power has been invested in the study of much larger systems, 10^4 - 10^6 atoms, in the appropriate environment, for example MD

studies of membrane-bound proteins embedded in model membrane environments are now commonplace [247]. As computing power increases, larger and more complex systems will become accessible to study using MD simulation, and on biologically relevant timescales.

Many of the challenges associated with the MD study of biomolecules pertain to the ability to study full systems, on biologically relevant timescales at the appropriate level of detail.

Recent advances and developments in the field of biomolecular simulation include the modelling of a complete virus in full atomistic detail, the use of MD to study protein folding, and the application and development of coarse-grained models to study larger and more complex biological systems.

2.11.1 MD Simulations of the Complete Satellite Tobacco Mosaic Virus

In 2006, Freddolino and co-workers published an all-atom 13 ns MD simulation of an entire life form, the full satellite tobacco mosaic virus (STMV). Due to the immense size of a combined virus water system, simulation of such a system had previously been computationally unfeasible. The simulation of the full virion system comprised 1,066,628 atoms; 899,565 water atoms, 135,960 protein atoms, 30,330 nucleic acid atoms and 773 ions. MD simulations of 10 ns were additionally performed on the RNA core and isolated capsid of the virion. The simulations were carried out using the CHARM22 force field [193,249] and performed with the NAMD 2.5 program [250].

While the full virion and isolated RNA were both dynamically stable on the simulation timescale, the isolated capsid swiftly became unstable and imploded in the absence of the RNA. This is consistent with the fact that assembled SMTV capsids are not found in the absence of any RNA core. Stability of the complete virion was maintained despite the absence of potentially favourable interactions; 12 N-terminal residues of each capsid monomer are absent due to disorder in the crystal structure. This region is thought to form favourable charge-charge interactions with the RNA that would further stabilise the full virion. Magnesium counterions, added primarily to neutralise the negative charge around the RNA core, remained attached to the RNA during the simulation in accordance with the generally ac-

cepted idea regarding the behaviour of complexes of DNA/RNA and ions [251].

The stability of the full virion and RNA, and instability of the capsid, demonstrated during these simulations adds further weight to the hypothesis (also suggested by the experimental findings of Day *et al* [252], Larson and McPherson [253], and Kuznetsov *et al* [254]) that in SMTV assembly the protein capsid assembles around a partially formed RNA core [251].

2.11.2 MD Simulations and Markovian State Models: Folding of the Villin Headpiece

In 2006 Jayachandran and Pande reported the examination of the dynamics of the 36-residue villin headpiece using large-scale distributed computing simulation and Markovian state models (MSMs) [255]. Tens of thousands of independent MD trajectories each several tens of nanoseconds in length were produced, to total sampling of nearly 500 μs. The majority of the simulations were performed on a subset of the 200, 000 processors participating in the group's Folding@Home distributed computing project [256]. Simulations were performed using a modified version of GROMACS 3.1.4 [235,257] and the "AMBERGS", a Garcia-Sanbonmatsu modified version of AMBER94, force field was used [258,259].

The work of Jayachandran and Pande concerned 1) directly simulating folding trajectories of explicitly solvated villin starting from an unfolded conformation, 2) an assessment of the sensitivity of dynamics to given system perturbations, and 3) the application of an MSM built using the collected trajectories to propagate dynamics to times beyond those simulated, and compute the evolution of ensemble property distributions over long timescales. The produced MSM was also used to predict the structure of villin *de novo* [255].

This work demonstrated the use of massively parallel simulation and analysis tools to overcome the barrier to the computational study of protein folding; the inaccessibility of long time scales and ensemble statistics. The simulation of the villin protein in explicit solvent enabled the rates and examination of folding trajectories to be computed, and showed that accurately built MSMs can propagate ensemble data over long times for each model [255].

2.11.3 Coarse-Grained Models

A number of simplified methods have been proposed to reduce the gap between the feasible time scales of classical atomistic MD simulations and those of biologically relevant motions. Simulations carried out using these reduced representation models are computationally less expensive than the equivalent atomistic MD simulations [260].

Explicit molecular mechanical treatment of the biomolecule and the use of a continuum model for the solvent has been applied to evaluate the solvation free energy of biological membranes [261]. An alternate simplified method is the coarse-grained (CG) model, where groups of atoms are represented as single interaction sites, commonly referred to as 'beads'. The CG method was first applied to a biomolecule in 1975 when Levitt and Warshel used a two bead model to represent the globular protein BPTI [262]. In this work each protein residue was represented by a pair of beads; one bead was centered on the Cα atom, and the other represented the side chain atoms. A torsion potential acted about the Cα beads, and a Lennard-Jones-type potential acted between pairs of side chain beads. This model successfully described the correct refolding of the protein from the starting point of a denatured configuration [260]. CG approaches have also been successfully applied to describe phospholipid bilayers [263], and viral capsids [264].

CG models contain fewer degrees of freedom and use force fields that lead to smother potential energy surfaces. Representing groups of atoms as single interaction sites typically allows for the use of longer time steps [260]. The development of CG models requires the identification of the 'important' degrees of freedom, and then the determination of the equations for describing the evolution of the system, considering only these degrees of freedom. This approach to overcome the spatial and timescale limitations of traditional MD comes at the cost of reduced accuracy [264] and as such CG methods cannot be used to investigate the molecular basis of recognition. The area of a protein that is involved in molecular recognition is often small and the associated recognition processes highly localised. Methods that integrate atomistic and CG approaches have been developed; one example is the hybrid MM/CG approach. This multi-level biomolecular simulation approach treats the small biologically relevant region of the protein at level of detail afforded by classical MD, while the

remainder of the protein is treated at the CG level of detail. As in hybrid QM/MM methods, an interface region is located between the MM and CG regions. This region bridges the discontinuity between the full atom and reduced representation descriptions [265].

2.12 Computational Methods Used in this Thesis

All MD simulations were performed using the GROMACS 3.2.1 simulation suite of programs (www.gromacs.org) [235] and the GROMOS96 43a2 united atom force field (ffG43a2). Simulations were performed either on University of Warwick Centre for Scientific Computing Argus task farm or Cluster of Workstations (COW) between April 2005 and December 2006. The source code for GROMACS 3.2.1 was compiled on each machine by the author of this thesis. Each simulation was run on a single node of one of the above machines. The approximate run time for a 1 ns simulation was 180 hours or 7.5 days. Docking simulations were carried out using the AutoDock 3.0.5 program [220] with the Lamarckian genetic algorithm (LGA). Version 8.1 of Modeller was used to model the missing active site loop (A3) in PheA and for homology modelling of the second A domain of Coelichelin CchH2.

Chapter 3

Molecular dynamics simulations of the phenylalanine activating adenylation domain, PheA, from *Bacillus brevis*

3.1 Introduction

In this chapter the results of a molecular dynamics (MD) simulation study of the L-Phenylalanine activating gramicidin S synthetase (GrsA) A domain (PheA) from *Bacillus brevis*[62] are presented. Although the A domains have been studied extensively and various sequence substrate specificity prediction models developed, understanding of the mechanism of substrate selectivity and the dynamics of the protein is still relatively rudimentary.

The NRPS A domain specifically selects and activates the amino/hydroxyl acid substrate through a two step reaction. In the first half reaction, a highly reactive aminoacyl adenylate is formed by reaction with Mg-adenosine triphosphate (ATP) resulting in the release of pyrophosphate. In the second half reaction the A domain binds the phosphopantetheinyl (PPant) arm of the downstream domain, the Peptidyl Carrier Protein (PCP) domain.

The PheA protein chain is folded into two distinct domains, a large A_{core} domain (1–412 pdb:17–428) and a smaller A_{sub} domain (413–514 pdb:429–530), which can be further divided into three and two sub-domains respectively, as outlined in chapter 1 section 1.4.2. The active site is located at the junction of the two structural domains. The L-Phe substrate is bound in a pocket accessible from the concave surface of the A_{core} domain near where the three A_{core}l sub-domains intersect[62]. The ten residues that line the L-Phe substrate binding pocket (pdb numbering in brackets) are: Asp 219 (235), Ala 220 (236), Trp 223 (239), Thr 262 (278), Ile 283 (299), Ala 285 (301), Ala 306 (322), Ile 314 (330) and Cys 315 (331) contributed by the A_{core} sub-domain, and Lys 501 (517), by the A_{sub} domain[62]. The Mg^{2+} ion was positioned in the structure by the authors. As PheA was the first "in module" A domain structure to be determined it has been used as a model for all subsequent A domain structural studies.

As outlined in Chapter 1 in section , "domain alternation" has been proposed as a strategy exploited by members of the adenylate-forming superfamily to reconfigure the single active site of the enzyme to perform the two half reactions. Structures of members of the adenylate-forming superfamily have been determined in the presence of the first and second half-reaction ligands[52,53,63,91]. These structures revealed that these enzymes have one active

site where the reactions take place, however the conformation of the structures differed with respect to the orientation of the A_{core} domain relative to the A_{sub} domain[52].

While A domains have only been determined in the adenylate-forming conformation, similarities between members of the adenylate-forming superfamily suggest NRPS A domains may exploit a similar strategy of domain alternation to reconfigure the enzyme's single active site. Members of the Adenylate forming superfamily contain highly conserved motifs and adopt a conserved fold. Residues from the core 5 motif (A8 motif) are are highly conserved in the A domains. These residues are critical for binding of the pantetheine portion of Coenzyme A (CoA) in the second half-reaction structures of members of the Adenylate forming superfamily. Limited proteolysis studies of the tyrocidine synthetase 1 A domain (TycA)[75,113] have indicated intrinsic flexibility of the protein in the region linking the A_{core} and A_{sub} domain, the first Arginine residue of the A8 motif, which was reduced in the presence of the first half-reaction ligands.

One way to probe conformation and examine the interaction between proteins and ligands is by using computer simulation, especially Molecular Dynamics (MD), which can provide information at the molecular level that is complementary to experiment and can, therefore, further the understanding of a system. To date, no molecular simulation study of the A domains has been reported in the literature. The MD simulations reported in this chapter were designed to explore the dynamics of the PheA A domain. Of particular interest was to probe the dynamical behaviour of PheA in the presence and absence of the hydrolysed products of the first half reaction.

These simulations of PheA reveal motion of the A_{sub} domain relative to the A_{core} domain. The principal modes of motion have been determined for PheA in each simulation. In each apo simulation the principal motion, described by eigenvector 1, describes the A_{sub} domain twisting clockwise, and tilting to the right, towards the A_{core} domain, and away from the A3 motif loop. In each holo simulation the principal motion shows the tilting and rotation of the A_{sub} domain (PheA2-holo), or part of the A_{sub} domain (PheA1-holo) towards the A3 motif loop. This loop is thought to play a key role in stabilising the phosphate atoms of ATP for the first half reaction and assist the removal of the pyrophosphate molecule from

the enzyme active site following the adenylation reaction. This domain motion is more pronounced in the PheA2-holo simulation where the A3 motif loop exhibits less flexibility and residues Thr 190 from the A3 motif loop form strong interactions with the highly conserved key L-Phe binding pocket residues Asp 235.

The rotation of the A_{sub} domain in the PheA1 and PheA2-holo simulations results in increased exposure between the domains on the right side of the protein. The extreme conformations of this motion were overlayed with a representative structure of the second half-reaction conformation (acetyl-CoA synthetase (bAcS) from *Salmonella enterica* pdb 1PG4) to identify the PheA phosphopantetheinyl binding site, and with the modular NRPS structure from the SrfAC synthetase to indicate the positioning of the PCP domain. This overlay indicates the principal motion of the A_{sub} domain in each holo simulation widens an opening between the domains on the right side of PheA which the flexible PCP domain and phosphopanteinyl arm could utilise to access the enzymes active site. The interaction between Thr 190 from the A3 motif loop and Asp 235 may be required to maintain the opening between the A_{core} and A_{sub} domain through which the PPant arm may access the PheA active site, or this interaction may be an intermediate stabilising interaction required to facilitate further rotation of the A_{sub} domain.

3.2 Methods

3.2.1 Simulation System

The crystal structure of PheA was determined by the multiple isomorphous replacement method to 1.9 Å resolution[62]. There are two copies present in the structure (pdb 1AMU) which have very similar conformations; the root mean squared deviation (RMSD) in the position of the main chain atoms of residues 21 to 530 after superposition is 0.26 Å. While both copies span the same region of the PheA sequence, copy A has fewer missing residues and side chain atoms and for this reason was used a starting structure for the MD simulations. Missing amino acid side chain atoms were added to the PheA structure using the

automated rotamer library of Swiss PDB Viewer[266] which selects the most energetically favourable rotamer.

No interpretable density was present for the 16 N-terminal residues or the 33 C-terminal residues. As these residues are not considered part of the core A domain structure no attempt was made to include these residues in the modelling. The core 2 motif (motif A3 in the A domains) is the most highly conserved motif in the superfamily of Adenylate forming enzymes. In PheA, this motif is consists of residues 190-TSGTTGNPKG-199 which form a loop between β-strands 5 and 6 in subdomain A. No significant electron density was determined for residues 192-GTTGN-196 in either copy of the molecule which implies that these central residues of the loop have conformational flexibility. The position of the remaining residues of the loop, with respect the to AMP binding site, suggests that this loop interacts with the pyrophosphate group which is displaced when ATP is hydrolysed to AMP. As this loop is thought to be required for the correct positioning of the phosphate groups of ATP and for facilitating the removal of pyrophosphate, it is of considerable interest[62,90].

The residues in this loop region were built into the structure and optimised using Modeller version 8.10[190,267]. 100 models of this loop region were constructed. The stereochemical quality of each loop model was evaluated using Procheck[197,198] and the energy of the models were assessed using the Modeller statistical potential DOPE[201]. The loop which corresponded to the lowest energy conformation was selected and this structure was used to initiate the MD studies.

3.2.2 Simulation Setup

The apo state of PheA (PheA-apo) was obtained by removing the L-Phe, AMP and Mg^{2+} ligands from the pdb structure. The holo state was taken from the PDB file. The GROMOS96 43a2 united atom force field (ffG43a2) was used for all simulations. Hydrogen atoms were added to the protein and phenylalanine ligand using the GROMACS pdb2gmx routine. Amino acid side chains of arginine and lysine were protonated, aspartic acid and glutamic acid as unprotonated, and histidine as neutral. In the case of the histidine residues,

the hydrogen atom was added in an automated fashion by the GROMACS pdb2gmx routine based on the optimal hydrogen bonding conformation.

The N- and C-terminal residues of the proteins were modelled as protonated and unprotonated respectively. One alternate method to this would be to cap the ends of the protein with neutral groups. This was not done as the termini of the protein project away from its core structure. The location of the N- and C-terminal residues was monitored throughout the simulations to ensure they did not make contact with the protein which could introduce artifacts. The phosphate group in AMP was modelled as deprotonated with an overall charge of -2. Since the ffG43a2 force field does not contain parameters for AMP, parameterisation of the AMP molecule was performed following published protocols[268]. The full method is described in section 3.2.7. Distance restraints were not used to constrain the Mg^{2+} ion to its starting position during the simulations.

Two sets of simulations of the apo and holo state PheA structure were performed to provide additional sampling. These shall be referred to as PheA1-apo and PheA1-holo (set 1), and PheA2-apo and PheA2-holo (set 2). Each set was subjected to a different minimisation and equilibration protocol, outlined in section 3.2.3.

All simulations were performed in a truncated octahedral box, 770 nm^3, and the GROMACS genbox routine was used to solvate the systems. This routine fills the box with multiple translational images of a single configuration of 216 simple point charge SPC[269] water molecules, then removes these water molecules when the distance between any atom of the solute molecule (protein or protein-ligand complex) and any atom of the solvent molecule is less than the sum of the van der Waals radii of both atoms. To achieve overall neutrality of the system, 16 randomly selected water molecules were replaced with Na^+ ions using the genion GROMACS utility. The resulting system sizes are listed in table 3.1.

Simulation	PheA1-apo	PheA2-apo	PheA1-holo	PheA2-holo
Protein atoms	5213	5213	5213	5213
Counterions (Na^+)	16	16	16	16
Water molecules	22952	22958	22933	22949
Total atoms	74085	74103	74078	74126

Table 3.1: Summary of PheA-apo and -holo simulation systems.

After minimisation (see section 3.2.3) each set of simulations was simulated in the canonical ensemble (constant **n**umber of particles, **v**olume and **t**emperature (NVT)) with heavy atoms tethered to ensure relaxation of the solvent:

- Set 1 was subjected to a 250 ps NVT MD simulation in which an isotropic force constant of 1000 kJ/mol^{-1} nm^{-1} was applied to tether all non-hydrogen atoms followed by 250 ps NVT MD simulation in which an isotropic force constant of 500 kJ/mol^{-1} nm^{-1} was applied to tether all heavy atoms.
- Set 2 was subjected to a 250 ps NVT MD simulation in which an isotropic force constant of 1000 kJ/mol^{-1} nm^{-1} was applied to tether all heavy atoms.

Subsequent to this, an un-tethered production run of 11.5 ns in the isothermalisobaric ensemble (constant **n**umber of particles, **p**ressure and **t**emperature (NPT)) was performed.

3.2.3 Energy Minimisation

Energy minimisation was used to relieve steric conflicts generated during the simulation setup. The convergence criteria for energy minimisation, $g = 0$ +/- e, is when the gradient (g) reaches a value within e of 0. Unless otherwise specified minimisation was performed until either e reached 1000 kJ mol^{-1} nm^{-1}, or the specified number of steps had completed. During energy minimisation no bonds were constrained. Unless otherwise specified, when heavy atoms were tethered a harmonic potential with a force constant of 1000 kJ mol^{-1} nm^{-2} was used.

3.2.4 PheA1-apo and -holo Energy Minimisation Protocol

The PheA1-apo and PheA1-holo systems were subjected to up to 100 steps of steepest descent minimisation with all heavy atoms tethered to their original position. After the addition of solvent (water and counterions), up to 100 steps of steepest descents minimisation

was performed with all heavy atoms tethered to their original position. This was followed by up to 100 steps of unrestrained conjugant gradients minimization.

3.2.5 PheA2-apo and -holo Energy Minimisation Protocol

The PheA2-apo and PheA2-holo systems were subjected to up to 1500 steps of steepest descent minimisation, or minimisation until *e* reached 100 kJ mol^{-1} nm^{-1}. During the minimisation all heavy atoms were tethered to their original position. After the addition of water, each system was subjected to 200 steps of steepest descent minimisation with heavy atoms tethered, or minimisation until *e* reached 100 kJ mol^{-1} nm^{-1}. This was followed by up to 3000 steps of unrestrained conjugant gradients minimisation. After the addition of counterions, up to 1000 steps of steepest descents minimisation was performed with all heavy atoms tethered to their original position. This was followed by up to 200 steps of unrestrained conjugant gradients minimization.

3.2.6 Simulation Protocol

All MD simulations were performed using the GROMACS 3.2.1 simulation suite of programs (www.gromacs.org)[235]. All heavy atoms were subject to a force constant of 1000 kJ mol^{-1} nm^{-2} during tethered runs. The particle mesh Ewald (PME)[238,239] method was used for the treatment of long range electrostatics with a cut-off of 1 nm and a Fourier spacing of 0.12 nm. The van der Waals interactions were modelled using a 1 nm cut-off. The 11.5 ns simulations were performed in the NPT ensemble; constant **n**umber of particles, **p**ressure and **t**emperature. The temperature of the simulation was kept constant by separately coupling the protein and solvent (water plus Na^+ counterions), and where appropriate, the AMP, L-Phe and Mg^{+2} to a Berendsen[242] thermostat at 310 K using a coupling constant, τ_T, of 0.5 ps. The pressure of the system was kept constant by isotropic coupling of the entire system to a pressure bath[242] at 1 bar using a coupling constant, τ_P, of 1.0 ps and a compressibility of 4.5 10^{-5} bar^{-1}. A time step of 2 fs was employed for all simulations. The centre of mass motion of the entire system was removed every step to maintain

the effective simulation temperature at 310 K. Initial velocities were generated at 300 K. During the MD simulations all bonds within the system were restrained using the LINCS algorithm[230]. Configurations were written to the trajectory file at every time step.

All MD simulations were performed on University of Warwick Centre for Scientific Computing Argus task farm or Cluster of Workstations (COW) between April 2005 and December 2006. The source code for GROMACS 3.2.1 was compiled on each machine by the author of this thesis. Each simulation was run on a single node of one of the above machines and the approximate run time for a 1 ns simulation was 180 hours or 7.2 days.

3.2.7 Development of AMP Force Field Parameters

Hydrogen atoms were added to the AMP coordinates from the PheA crystal structure using QUANTA. The AMP coordinates were then submitted to the Dundee PRODRG2 Server[270] to define the GROMOS skeleton topology file. The output force field parameters from PRODRG2 were converted manually from the GROMOS87 into the GROMOS96 format. The values provided as ideal bond lengths, angles and dihedrals (improper and proper) by PRODRG2 were compared to those of ATP in the ffG43a2 residue topology file and modified accordingly. The pairs list was modified by examining the ATP ffG43a2 exclusions list.

Explicit hydrogen atoms were defined on the adenine ring consistent with the development of GROMOS96 parameters for flavin adenine dinucleotide (FAD) by Berg and coworkers[271]. As suggested in the literature[62], the oxygen atoms in the phosphate group of the AMP molecule were defined as unprotonated. Charges for the force field parameters of the PO_4^{-2} region of the AMP molecule were derived using *ab initio* QM calculations performed using Gaussian 98. These calculations were performed at a number of different levels of theory for comparison. In line with previous studies[268] the starting charges for the phosphate region were taken from calculations performed at the HF/6-31G* level of theory. These charges were subsequently refined and scaled as has been employed when deriving previous GROMOS96 force field parameters for other ligands[272]. The charges derived from

the ab initio calculation and scaled for use with the GROMO96 force field are listed in table 3.2.

Atom/Type	**O1/OM**	**O2/OM**	**O3/OM**	**P/P**	**O5*/OM**
Calculated	-0.999518	-1.015533	-0.997015	1.510475	-0.594859
Used	-0.990	-0.990	-0.990	1.490	-0.520

Table 3.2: AMP PO_4^{-2} partial charges, calculated using HF/6-31G* and scaled according to GROMOS force field conventions.

Using these initial force field parameters an AMP molecule was simulated under periodic boundary conditions in a truncated octahedral box of volume 26.92 nm^3. The system was solvated using the genbox GROMACS routine. Of the 861 added water molecules, two were randomly selected and replaced with Na^+ ions to achieve overall neutrality. This system was then subjected to energy minimisation using steepest descent and then conjugant gradient methods to a convergence tolerance of 0.001 kJ mol^{-1}. A 250 ps NVT MD simulation with the AMP heavy atoms tethered to their original positions with an isotropic force constant of 1000 kJ/mol^{-1} nm^{-2} was carried out. Finally, the last configuration from the restrained MD was used to initiate a NPT production run of 1 ns using the simulation protocol detailed previously in section 3.2.6.

The RMSD of the AMP molecule in this simulation showed it to be structurally stable on the timescale of the simulation. No significant conformational changes of the AMP structure were observed during the 1 ns simulation. These force field parameters were therefore used in the simulations presented in this thesis. The complete topology file and RMSD of the AMP molecule can be seen in table 7.2 and figure 7.8 of appendix 7.1.2 respectively.

3.2.8 MD Simulation Analysis Methods

Trajectories were analysed using GROMACS routines, custom written VMD tcl scripts, and the DSSP program[273]. Essential degrees of freedom of PheA were extracted from the production run trajectories according to principal components analysis (PCA) or the essential dynamics (ED) method, performed using the GROMACS suite, and analysed using the DynDom server[274,275].

The RMSD and root mean square fluctuations (RMSF) of the protein Cα atoms were analysed using GROMACS routines. Ligands coordinated to the magnesium ion (atoms within 0.36 nm) were identified using VMD tcl scripts written by the author. Hydrogen bonding interactions were analysed using custom written VMD tcl scripts; where a bond was considered to be present within a cut off of 60° and 0.36 nm. Secondary structural analysis as a function of time was assessed using the DSSP program[273].

Principal Components Analysis

The dynamical properties of the PheA protein in each simulation were quantitatively characterised using PCA[276]. This method describes the positional fluctuations of individual atoms in terms of a set of mutually linearly independent collective fluctuations. A PCA is performed by first constructing a covariance matrix. Before the covariance matrix is constructed, rotational and translational degrees of freedom are removed from the trajectory by performing a least squares fit to a reference structure; the subset of atoms used for the fitting was the backbone atoms[277,278].

Once constructed, the covariance matrix is diagonalized to produce a set of eigenvectors and eigenvalues. Motion along a single eigenvector corresponds to concerted fluctuations of atoms and the eigenvalues represent the total mean square fluctuation of the system along the corresponding eigenvectors. Eigenvectors are sorted by size and the first has the largest eigenvalue. In proteins, only a few eigenvectors have large eigenvalues and these are described as "essential eigenvectors". It is assumed that for proteins the most biologically significant motions would be described by these few essential eigenvectors[277,278].

To further analyse the nature of the collective motions of PheA in each system the trajectories were projected onto the respective first ten eigenvectors to reveal the sampling along these vectors. The extreme projections of the trajectories along the first ten eigenvectors were obtained. These structures were processed using the DynDom program[274,275] which compares pairs of structures to determine the relative motions of domains, residues that act as hinges between the domains, and the quantitative nature and volume of interdomain

motion with respect to a reference axis. Identification of domains is done by performing a whole protein best fit of the two structures and determining rotation vectors of short main-chain segments. Then using an algorithm to identify clusters of rotation vectors where each individual cluster of residues segments forms a possible dynamic domain. Residues proposed as hinges between the identified domains are used to define a hinge axis and the motion of the dynamic domain is quantitatively described in reference to this axis[274,275].

3.3 Results and Discussion

Analysis of the PheA-apo and -holo simulations is presented in this section. Results for the entire production run of 11.5 ns have been analysed and are presented. The implications of these observations are discussed in detail in relation to the biological relevance and simulation methods used.

Simulations presented in the later chapters of this thesis will be compared with these initial simulations of the PheA crystal structure in the apo and holo states. Observations are therefore discussed in detail to provide a benchmark against which the further simulations will be compared.

3.3.1 Global Structural Stability

Information regarding the conformational stability of the protein on the timescale of the simulation is provided by comparing the structural drift (Root Mean Square Deviation - RMSD) of the protein from its corresponding starting structures after least-squares fitting (rigid body rotation and translation).

In figure 3.1 the all atom Cα RMSDs of PheA from the corresponding starting structures are shown as a function of time for each of the four simulations. RMSDs from the initial structure are stable during the PheA-apo2 simulation. The RMSD rises during the first nanosecond to an initial plateau of $\sim$0.22 nm which is maintained until $\sim$5 ns when, be-

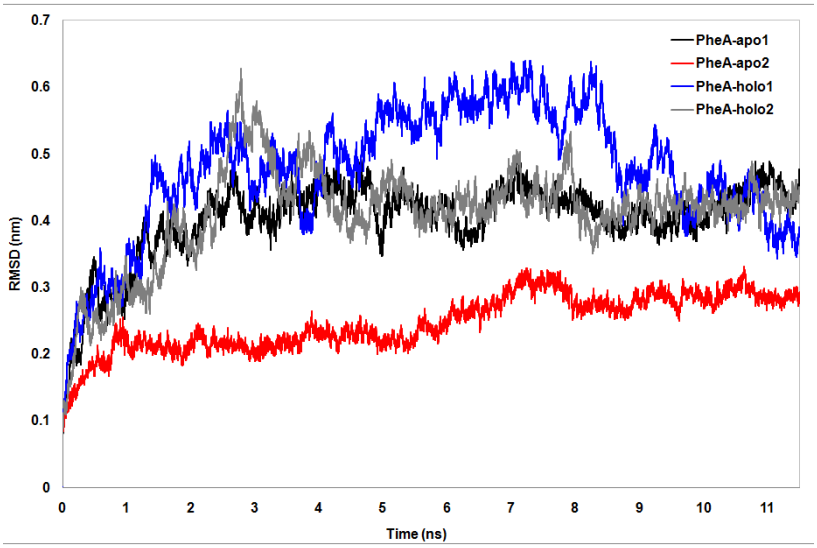


Figure 3.1: **All atom Cα RMSDs PheA-apo and -holo simulation.** The conformational drift of PheA measured as Cα atom root mean square deviation (RMSD) from the starting structure for the PheA1-apo (black), PheA2-apo (red), PheA1-holo (blue) and PheA2-holo (grey) simulations.

tween 5 and 7.8 ns the all Cα RMSD rises to peak at ∼0.33 nm. A plateau is reached at 7.8 ns when the RMSD attains a value of approximately 0.3 nm. The all Cα atoms RMSDs for the PheA1-apo, PheA1-holo and PheA2-holo simulations increase from the beginning of the simulation and attain high values of 0.5, 0.65 and 0.63 nm respectively. While the RMSD of the Cα atoms in the PheA1-apo simulation is higher than that in the PheA2-apo simulation, the evolution of the RMSD as a function of time is similar. In the PheA2-apo simulation the RMSD rises during the first nanosecond to ∼0.35 nm. It then rises to ∼0.49 nm at 4.6 ns after which it fluctuates between 0.35 and 0.5 nm for the remainder of the simulation.

The all atom Cα RMSD as a function of time for the PheA1-holo simulation resembles an arc. The RMSD rises steadily during the first 7.3 ns to peak at ∼6.4 nm. Between 8 ns and 11.5 ns the RMSD decreases to 0.37 nm. The all atom Cα RMSD from the PheA2-holo simulation exhibits a similar trend, although on a different timescale. The all atom Cα RMSD peaks at ∼0.63 nm at 3 ns, fluctuating between 0.54 and 0.36 nm until 9 ns, when it settles to fluctuate about ∼0.42 nm.

The high values of the all Cα atoms RMSDs for the PheA1-apo, PheA1-holo and PheA2-holo simulations, particularly as compared to the PheA2-apo simulation, indicates there

are significant structural changes of PheA in these simulations. To understand whether a particular region of PheA is contributing to the high RMSD values in these simulations or if it is as a result of global structural distortions, the protein was decomposed into its component domains (A_{core} and A_{sub}). To remove the contribution of the more flexible loops and linker region from the core secondary structural elements which are expected to be more stable, the individual domains were also split into the secondary structural elements (helices and sheets), loops and linker region.

The A_{core} and A_{sub} domain were defined as residues 24 to 413, and 414 to 514 respectively. The residues forming secondary structural elements and loops were identified using a combination of the authors[62] annotation of the PheA structure, visualisation of the PheA starting structure and the output from the DSSP program. The N- and C-terminal linker regions were defined as containing residues 1 to 23 (pdb: 17 to 39) and 503 to 514 (pdb: 519 to 530) respectively. The RMSD of each of these regions, after least-squares fitting to the relevant region of the initial starting structure, was calculated for each simulation.

The Cα atom RMSDs of these regions of PheA are shown in figures 3.2, 3.3, 3.4 and 3.5 for the PheA1-apo, PheA2-apo, PheA1-holo and PheA2-holo simulations respectively.

The upper graph in figure 3.2 shows the time dependent RMSD of all the Cα atoms (black), and the A_{core} (red) and A_{sub} (blue) subdomains Cα atoms in the PheA1-apo state simulation. RMSDs for the A_{core} domain from the initial structure are stable throughout the simulation and converged to values of $\sim$0.27 nm, showing a close resemblance to the starting structure. RMSD of the A_{sub} domain rises to reach an initial plateau of $\sim$0.25 nm between 2 and 5 ns, and a later plateau of $\sim$0.33 nm at 6 ns which is maintained for the remainder of the simulation. While the A_{sub} domain is stable on the timescale of the simulation, the larger A_{core} domain is more stable. The large RMSD of the all the Cα atoms in not accounted for by the RMSDs of the individual domains suggesting that there may by motion of the domains relative to one another.

The middle and lower graphs of figure 3.2 show the RMSD of the Cα atoms of the individual components of the A_{core} and A_{sub} domains of PheA1-apo, respectively. From

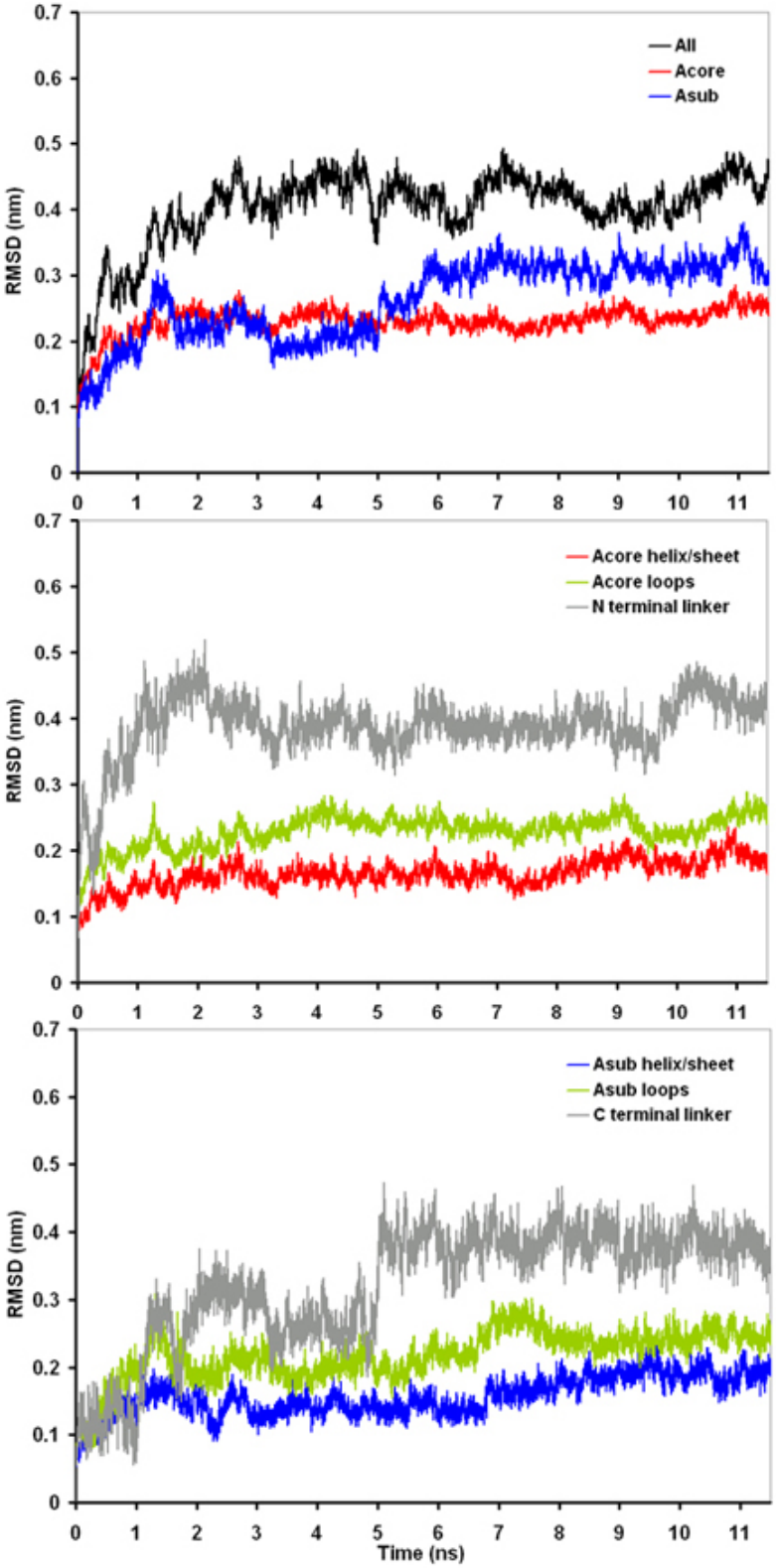


Figure 3.2: **RMSD PheA1-apo simulation.** The conformational drift of PheA1-apo, measured as Cα atom root mean square deviation (RMSD) from the starting structure. RMSD vs. time is shown for the entire protein (black), the A_{core} domain (red) and A_{sub} domain (blue), in the upper graph. The RMSD of the linker (grey), secondary structural elements - helices and sheets (red), and loop regions (green) of the A_{core} domain in shown in the middle graph, and the RMSD of the equivalent regions of the A_{sub} domain in the lower graph.

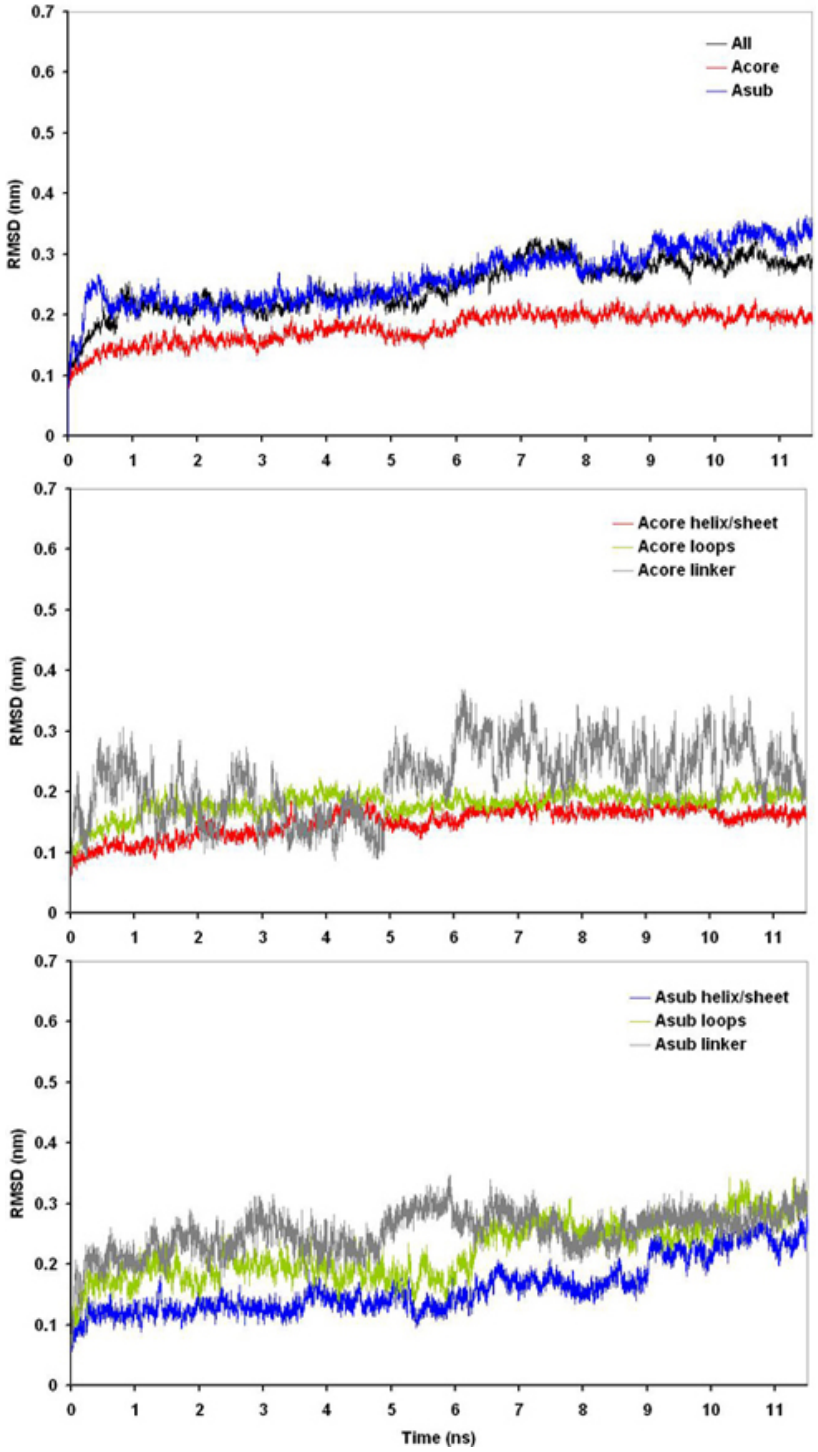


Figure 3.3: **RMSD PheA2-apo simulation.** The conformational drift of PheA1-apo, measured as Cα atom root mean square deviation (RMSD) from the starting structure. RMSD vs. time is shown for the entire protein (black), the A_{core} domain (red) and A_{sub} domain (blue), in the upper graph. The RMSD of the linker (grey), secondary structural elements - helices and sheets (red), and loop regions (green) of the A_{core} domain in shown in the middle graph, and the RMSD of the equivalent regions of the A_{sub} domain in the lower graph.

the RMSDs of these regions it is clear that the largest deviation is observed in the N- and C-terminal linker regions.

Both the Cα RMSD of the secondary structural elements (the α-helices and β-strands) of the A_{core} and A_{sub} domains display a comparable level of stability on the timescale of the simulation. The RMSD of the secondary structural elements of the A_{core} domain reaches a plateau of $\sim$0.17 nm at 2 ns rising to $\sim$ 0.2 nm from 9 ns to the end of the simulation. The Cα RMSD of the secondary structural elements of the A_{sub} domain reach an initial plateau of $\sim$0.15 nm between 1 and 6.8 ns, rising to fluctuate between 0.2 and 0.3 nm, ending the simulation at 0.25 nm.

The loop regions of each domain exhibit greater structural stability on the timescale of the simulation than the N- and C-terminal linker regions, as one may expect given they are largely short regions of sequence anchored by the secondary structure regions. Although of larger magnitude, the overall trend of the RMSD for the loop regions of each domain is similar to that of the RMSD of the secondary structural elements in the respective domain.

The overall pattern of the RMSD of the component portions of each domain of PheA in the PheA1-apo simulation is similar. RMSD of the secondary structural elements of each domain shows comparable structural stability of the A_{core} and A_{sub} domains, with the N- and C-terminal linkers exhibiting the greatest structural drift.

The RMSD of the component regions of PheA in the PheA2-apo state simulation are shown in figure 3.3. The upper graph shows the time dependent RMSDs of the Cα atoms, A_{core} and A_{sub} domain. The middle and lower graphs show the time dependent RMSDs of the Cα atoms of the individual components of the A_{core} and A_{sub} domains from the equivalent region of the starting structure.

The RMSDs of the Cα atoms and A_{sub} domain for the PheA2-apo simulation are larger and the overall trend of each is distinctly different from those from the PheA1-apo simulation. As in PheA1-apo, on the timescale of the simulation the A_{core} domain of PheA from Phe2-apo exhibits the greatest structural stability. The A_{sub} domain exhibits slightly greater structural drift in the PheA2-apo simulation, as compared to the PheA1-apo simulation,

fluctuating on the time scale of the simulation from ∼0.24 nm at 1ns to ∼0.35 nm at 11.5 ns.

The RMSD of the A_{core} domain secondary structural elements from PheA2-apo is comparable in trend, however attains slightly lower values, as compared with the PheA1-apo simulation. The RMSD of the A_{sub} domain is comparable in the PheA1 and Phe2-apo simulations during the first 10 ns, exhibiting slightly greater structural drift towards the end of the PheA2-apo simulation.

The RMSD of the A_{core} domain loop regions attain higher values in the PheA1-apo simulation than PheA2-apo simulation, indicating slighter greater structural drift of this region, as compared with the starting structure, in the PheA2-apo simulation. As in the PheA1-apo simulation the overall trend of the RMSD for the loop regions of each domain is similar to that of the RMSD of the secondary structural elements in the respective domain.

The N-terminal and C-terminal linker region RMSDs in the PheA2-apo simulation fluctuate between 0.10 and 0.37 nm, and 0.18 nm and 0.33 nm on the timescale of the simulation respectively, attaining lower values than in the PheA1-apo simulation. The loop regions of the A_{core} domain exhibit greater structural stability on the timescale of the simulation than the N-terminal linker region which exhibits the largest structural drift of the components in the domain. The A_{sub} domain loop regions exhibit greater structural stability than the C-terminal linker region for the first half of the simulation. During the last 5 ns of the simulation the RMSD of the A_{sub} domain loop regions rises to between 0.25 and 0.35 nm, intermittently attaining a higher RMSD value than the C-terminal linker.

The magnitude of the RMSD of the three components of the A_{sub} domain relative to one another is similar to that of the A_{core} domain in the PheA2-apo simulation.

The secondary structural elements of the A_{core} domain exhibit greater structural stability in the PheA2-apo simulation than in the PheA1-apo simulation. The stability of the secondary structural elements of the A_{sub} domain is comparable across the two apo simulations with the exception of the final 1.5 ns of the PheA2-apo simulation where the RMSD of the A_{sub} domain increases by 0.4 nm. There is greater structural stability of the A_{core} domain loops

in the PheA2-apo simulation than the PheA1-apo simulation and comparable stability of the A_{sub} domain loops. The N- and C-terminal linker regions exhibit greater structural drift in the PheA1-apo simulation.

The RMSD of the component regions of PheA from the PheA1-holo and PheA2-holo simulations are shown in figures 3.4 and 3.5 respectively. As for the apo simulations, the upper graph shows the time dependent RMSDs of all the PheA Cα atoms, A_{core} and A_{sub} domain. The middle and lower graphs show the time dependent RMSDs of the Cα atoms of the individual components of the A_{core} and A_{sub} domains from the equivalent region of the starting structure.

The RMSD of the Cα atoms of the A_{core} and A_{sub} domain, and the individual domain components are remarkably similar for the PheA1- and PheA2-holo simulations. These results will, therefore, be discussed together. The upper graph of figures 3.4 and 3.5 show the RMSD of all the Cα atoms, the A_{core} and A_{sub} Cα atoms for the PheA1-holo and PheA2-holo simulations respectively. As previously outlined the overall trend of the all atom Cα RMSD resembles an arc; reaching a peak of $\sim$6.4 nm at 7.3 ns. The all atom Cα RMSD of PheA in the PheA2-holo simulation also resembles as arc; reaching a peak of $\sim$0.63 nm at 3 ns.

The trend of the A_{core} and A_{sub} domain Cα RMSDs are similar in the PheA-holo simulations. As in the apo simulations the A_{core} domain is more stable on the timescale of the simulation that the A_{sub} domain. In the PheA1-holo simulation the A_{sub} Cα atom RMSD fluctuates about $\sim$0.17 nm until 3.3 ns, after this point it increases to a peak of $\sim$0.37 nm over 4.7 ns, fluctuating about 0.3 nm for the remainder of the simulation. In the PheA2-holo simulation the A_{sub} Cα RMSD gradually increases over the first 5.2 ns to peak at $\sim$0.36 nm, after which time it fluctuates about an average of $\sim$0.3 nm.

As in the PheA1-apo simulation the large RMSD of the all the Cα atoms in these simulations in not accounted for by the RMSDs of the individual domains suggesting that there may by motion of the domains relative to one another.

The middle and lower graphs of figure 3.4 and 3.5 show the Cα RMSD of the individual

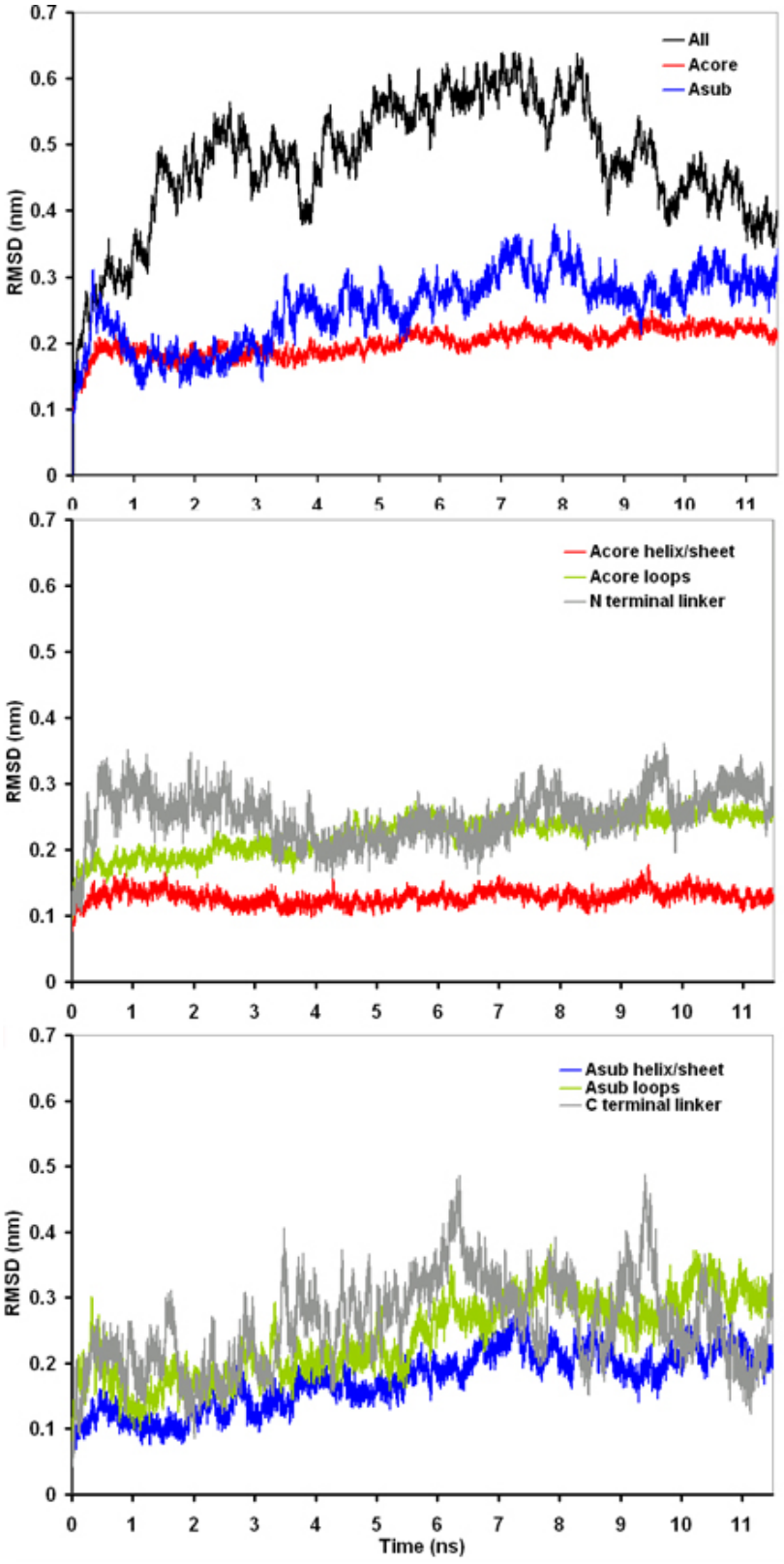


Figure 3.4: **RMSD PheA1-holo simulation.** The conformational drift of PheA1-apo, measured as Cα atom root mean square deviation (RMSD) from the starting structure. RMSD vs. time is shown for the entire protein (black), the A_{core} domain (red) and A_{sub} domain (blue), in the upper graph. The RMSD of the linker (grey), secondary structural elements - helices and sheets (red), and loop regions (green) of the A_{core} domain in shown in the middle graph, and the RMSD of the equivalent regions of the A_{sub} domain in the lower graph.

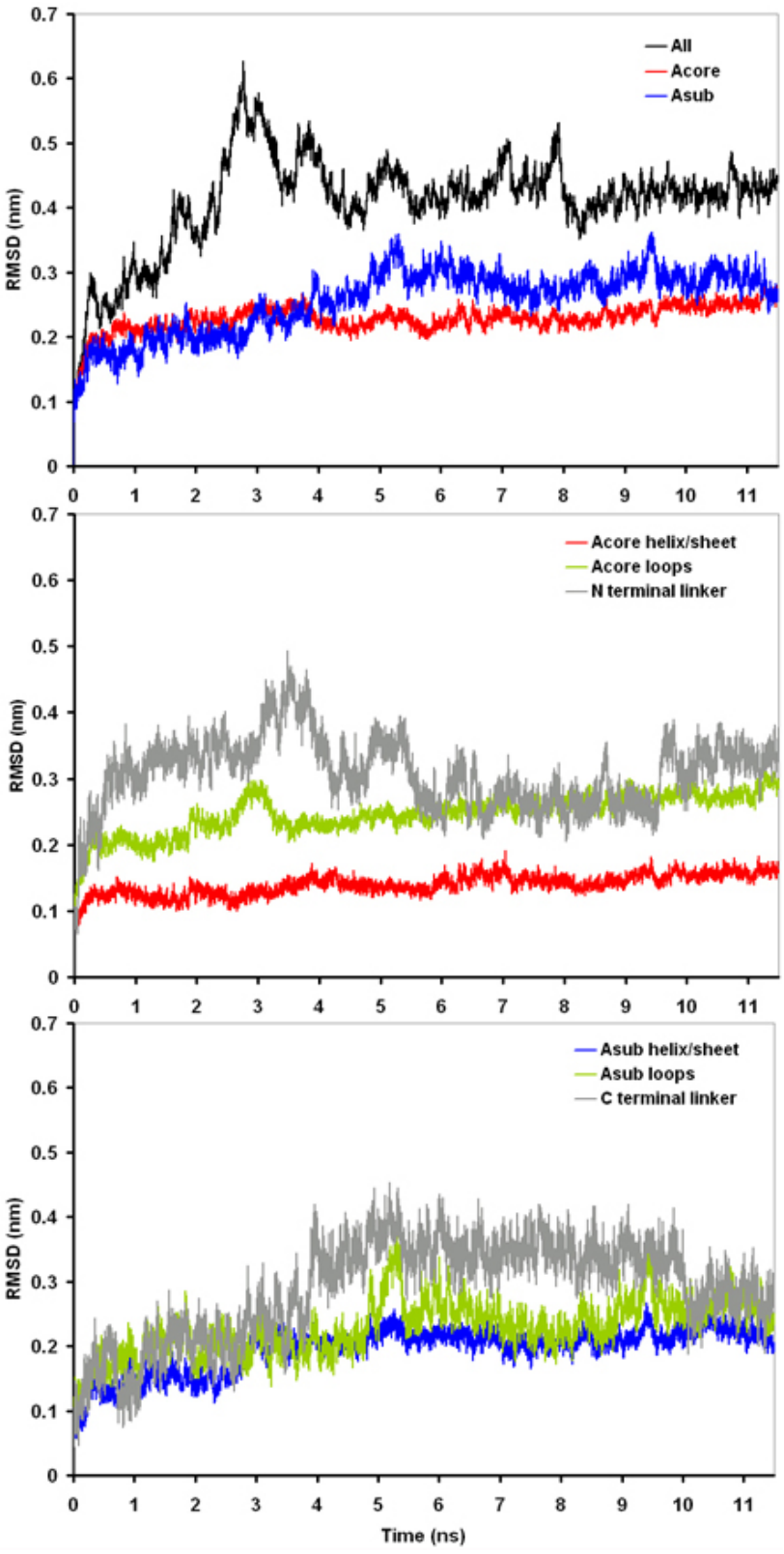


Figure 3.5: **RMSD PheA2-holo simulation.** The conformational drift of PheA1-apo, measured as Cα atom root mean square deviation (RMSD) from the starting structure. RMSD vs. time is shown for the entire protein (black), the A_{core} domain (red) and A_{sub} domain (blue), in the upper graph. The RMSD of the linker (grey), secondary structural elements - helices and sheets (red), and loop regions (green) of the A_{core} domain in shown in the middle graph, and the RMSD of the equivalent regions of the A_{sub} domain in the lower graph.

components of the A_{core} and A_{sub} domains of PheA1-holo and PheA2-holo simulations, respectively. In both simulations the secondary structural elements of the A_{core} domain are stable on the timescale of the simulation. As the all Cα atom RMSD of the A_{sub} domain indicates, the A_{sub} domain structural elements exhibit greater structural drift on the timescale of the simulation than the A_{core} domain secondary structural elements. In PheA1-holo the RMSD of the Cα atoms of the secondary structural elements of the A_{sub} domain rises during the first 7.5 ns of the simulation to peak at $\sim$0.27 nm. For the remainder of the simulation the RMSD of this region fluctuates about 0.22 nm. The overall trend of the RMSD for the equivalent region of PheA2-holo is similar to that seen in the PheA1-holo simulation. The RMSD of this region increases during the first 6 ns of the simulation to peak at 0.25 nm and for the remainder of the simulation the RMSD fluctuates about 0.22 nm.

In the PheA1-holo simulation, the RMSD of the Cα atoms of the A_{core} loops gradually increases on the simulation timescale, starting at $\sim$0.17 and ending at $\sim$0.26 nm. A similar pattern is observed in the PheA2-holo simulation with the RMSD of this region rising initially to $\sim$0.2 nm and then steadily increasing to reach $\sim$0.28 nm by the end of the simulation. The overall trend of the RMSD of the A_{sub} loop Cα atoms in both holo simulations is similar with the RMSD increasing throughout the simulation. In the PheA1 the RMSD of this region rises throughout the simulation to $\sim$0.3 nm by 11.5 ns, and in PheA2 the RMSD of this region is $\sim$0.27 nm at 11.5 ns. This behaviour is comparable with that observed in the apo simulations.

The N- and C-terminal linker regions exhibit greater structural drift that the other elements of the A_{core} and A_{sub} domains in both holo simulations. The N-terminal linker exhibits greater conformational stability in the PheA1-holo simulations than the PheA1-apo simulations, this region has an average RMSD of 0.25 nm and 0.39 nm in these simulations respectively. Conversely, the N-terminal linker exhibits greater conformational stability in the PheA2-apo simulation than the PheA2-holo simulation.

The magnitude of the RMSD for the C-terminal linker Cα atoms is comparable in the holo simulations however the trend is different. The RMSD of this region in the PheA1-holo

simulation increases gradually for the first 5.3 ns to reach $\sim$ 0.43 nm, it then fluctuates at an average of $\sim$0.35 nm until 10 ns and then falls to an average of $\sim$0.28 nm at the end of the simulation. The RMSD of this region in the PheA1-holo simulation fluctuates throughout the simulation between 0.15 and 0.45 nm.

Summary

The RMSD of the Cα atoms of the entire individual A_{core} and A_{sub} domains is very similar in all simulations, with the A_{core} domain exhibiting the greatest conformational stability. Decomposing the domains into the core secondary structural elements, loops, and linker regions reveals greater conformational stability of the secondary structural elements of the A_{core} domain in the holo simulations than in the apo simulations. Conversely, greater structural drift was observed in the A_{sub} domain of the holo simulations as compared with the apo simulations.

The high RMSD value of the all Cα atoms of PheA in the PheA1-apo, PheA1-holo and PheA2-holo simulations coupled with the identification that in each of these simulations the A_{core} and A_{sub} domains are structurally stable suggests that there may be motion of the A_{core} and A_{sub} domains relative to one another. Such motion would be magnified during the fit of the entire protein to itself for the calculation of the all Cα atoms RMSD.

3.3.2 Radius of Gyration

The radius of gyration (Rg) of a protein provides an indication of the compactness of the structure. The Rg was calculated for PheA in each simulation and the average Rg is shown in table 3.3.2. In each simulation the average Rg for PheA has decreased slightly from the starting structure indicating the structures may have become slightly more compact. This is slightly more pronounced in the apo simulations than the holo simulations. Plotting the Rg of PheA from each simulation against time revealed no significant differences between the Rg evolution of PheA in each simulation.

	Starting	PheA1-apo	PheA2-apo	PheA1-holo	PheA2-holo
RG (nm)	2.38	2.33 (0.03)	2.34 (0.01)	2.35 (0.02)	2.36 (0.03)

Table 3.3: **Average values of radius of gyration** (Rg) for the apo and holo PheA simulations. Standard deviation in parentheses.

3.3.3 Secondary Structure

In table 3.3.3 the average secondary structure content in PheA for each of the simulations, and in the starting structure, according to DSSP classification, is reported. Visual plots of the evolution of the secondary structure content versus time for each simulation are included on the accompanying CD.

These analyses revealed that all β-sheets and α-helices within the core regions of PheA are well maintained throughout each entire simulation. While stable, a number of key regions of the structure, involved in either ligand binding or comprised of residues from the conserved motifs, exhibit notable behaviour.

The first such region is the stretch of eleven residues (407–417, pdb: 423–434) linking the A_{core} domain to the A_{sub} domain. This region contains the highly conserved L-Asp residue (414, pdb: 430), referred to as the hinge residue. Residues 407–417 are annotated in the crystal structure by Conti *et al.*[62], as two small distinct strands (although they are referred to as strand C5). In the PheA1-apo, PheA1-holo, and PheA2-holo simulations, where the all Cα atom RMSD of PheA indicates motion between the A_{core} and A_{sub} domain, these strands merge to form one long strand that persists on the timescale of the simulation. In the PheA2-apo simulation however, the two smaller strands largely remain distinct, and a single strand is only formed intermittently.

In the PheA structure the invariant L-Lys residue (501, pdb: 517) that forms electrostatic interactions with the L-Phe substrate is located on the long loop (A10 motif K loop) that projects from the A_{sub} domain into the binding pocket. In all four simulations, an antiparallel β-sheet is intermittently formed by the residues either side of this lysine residue.

In the PheA1-apo simulation this region predominantly adopts the β-sheet conformation

Secondary Structure	Starting	PheA1-apo	PheA2-apo	PheA1-holo	PheA2-holo
Coil	105.0	110.66 (5.2)	111.53 (5.48)	109.96 (5.83)	113.11 (6.16)
B-Sheet	135.0	125.76 (6.23)	123.88 (5.72)	128.40 (6.19)	124.93 (6.32)
B-Bridge	4.0	4.86 (2.35)	6.63 (2.10)	3.86 (6.28)	6.42 (2.05)
Bend	50.0	66.21 (6.83)	66.30 (5.22)	66.52 (6.28)	61.83 (5.78)
Turn	69.0	57.52 (6.24)	53.25 (5.79)	52.32 (6.10)	60.62 (7.16)
A-Helix	136.0	138.56 (5.59)	147.92 (3.77)	147.42 (3.69)	139.55 (5.28)
5-Helix	0	4.88 (3.88)	0.004 (0.20)	0.004 (0.21)	2.15 (2.78)
3-Helix	15.0	5.56 (3.28)	4.49 (3.00)	5.52 (3.02)	5.38 (3.17)

Table 3.4: **Average secondary structure distribution in PheA apo and holo simulations** and in the starting structure. Table shows average number of residues per secondary structural element as measured throughout entire simulation. Standard deviations are in parentheses.

between 1.6 and 5.95 ns and this sheet is seen intermittently between 6 ns and 10.4ns. In the PheA2-apo simulation this sheet is first formed at 0.7 ns and is frequently present between 4.9 ns and 7.9 ns, after which time this region adopts a bend structure. In the PheA1-holo simulation this region is initially dominated by a turn/bend/turn structure until 1.5 ns where the anti-parallel sheet is observed intermittently until 6.8 ns. After 6.8 ns this region forms a bend until 8.55 ns, after which time the anti-parallel beta sheet is re-established and remains until the end of the simulation. Quite a different pattern of secondary structure formation is seen in the PheA2-holo simulation. In this simulation these residues only adopt the sheet structure between 3.2 and 3.4 ns. During the rest of the simulation these residues primarily adopt a turn structure.

The secondary structure of three other distinct regions of PheA fluctuate on the time scale of the simulation. These regions are all located on the external faces of the protein, away from both the active site and the domain linker, and in the starting structure are either long unstructured regions or are in the N- or C-terminal linker regions.

Residues 4 –13 (pdb: 20–29) form α-helix H1, see figure 1.10. Based on the authors[62] annotation of the PheA structure this helix is not considered part of the core A domain structure. Instability in this region may be as a result of the missing N-terminal 16 residues of the protein which may stabilise this helix. Secondary structure prediction suggests six, ILIHAQ, of the 16 missing residues (MVNSSKS ILIHAQNKN) may form a β-sheet. Helix H1 is stable throughout the PheA1-holo and PheA2-apo simulations. In the PheA2-holo simulation and PheA1-apo simulation helix H1 exhibits increasing instability at the C-terminal end during the initial stages of simulation. A π-helix is formed from these residues and observed intermittently during the PheA2-holo simulation and consistently throughout the PheA1-apo simulation. These findings are consistent with the RMSD of the N-terminal linker region, which are lower in value for the PheA1-holo and PheA2-apo simulations, than the PheA2-holo and PheA1-apo simulations.

A long unstructured region of sequence comprising residues 148–166 (pdb: 164–182) links strands A4 to A5. These residues are located on the exterior face of the protein at the base of subdomain A of the A_{core} domain and they are positioned across helix H3 which is

formed by the A1 motif residues; this motif is thought to be conserved as it contributes to the overall stability of the A domain structure. Residues 148-166 form a very short α-helix in simulations PheA2-apo, PheA1-holo and PheA2-holo, and intermittently in PheA1-apo.

Helix H5, which is formed by residues 374–380 (pdb: 390–396), is situated at the interface of subdomains B and C of the A_{core} domain. This helix remains stable on the timescale of the PheA1-apo, PheA2-apo, and PheA1-holo simulations. In the PheA2-holo simulation the helix is stable until 3.3 ns, where for the remainder of the simulation these residues predominantly form a π-helix.

3.3.4 Structural Flexibility

A measure of the relative flexibility of different regions of PheA throughout the simulations can be obtained by calculating the residue-by-residue fluctuations (Root Mean Squared Fluctuations - RMSFs) of the simulated structures relative to the average structure. Figure 3.6 shows the comparison of the Cα atom RMSF from the PheA1-apo and PheA2-apo simulations (upper left graph), PheA1-apo and PheA2-holo simulations (lower left graph), PheA1-apo and PheA1-holo simulations (upper right graph), and PheA2-holo and PheA2-apo simulations (lower right graph).

In each of the simulations the residues of the A_{sub} domain of PheA display higher flexibility than any region of the A_{core} domain. This correlates with the observations from the RMSD that greater structure drift occurs in the A_{sub} domain as compared to the A_{core} domain. The PheA1-holo simulation displays the most flexibility in this region. Qualitatively flexibility of A_{sub} domain is similar in PheA1-apo to PheA1-holo, however, it is of slightly lower magnitude. The flexibility of the A_{sub} domain in PheA2-holo is very similar to that from PheA1-holo. Notable regions that exhibit lower flexibility in the PheA2-holo simulation are:

- the loops immediately preceding the first β-sheet of subdomain E
- the loops following the second and third E subdomain β-sheets

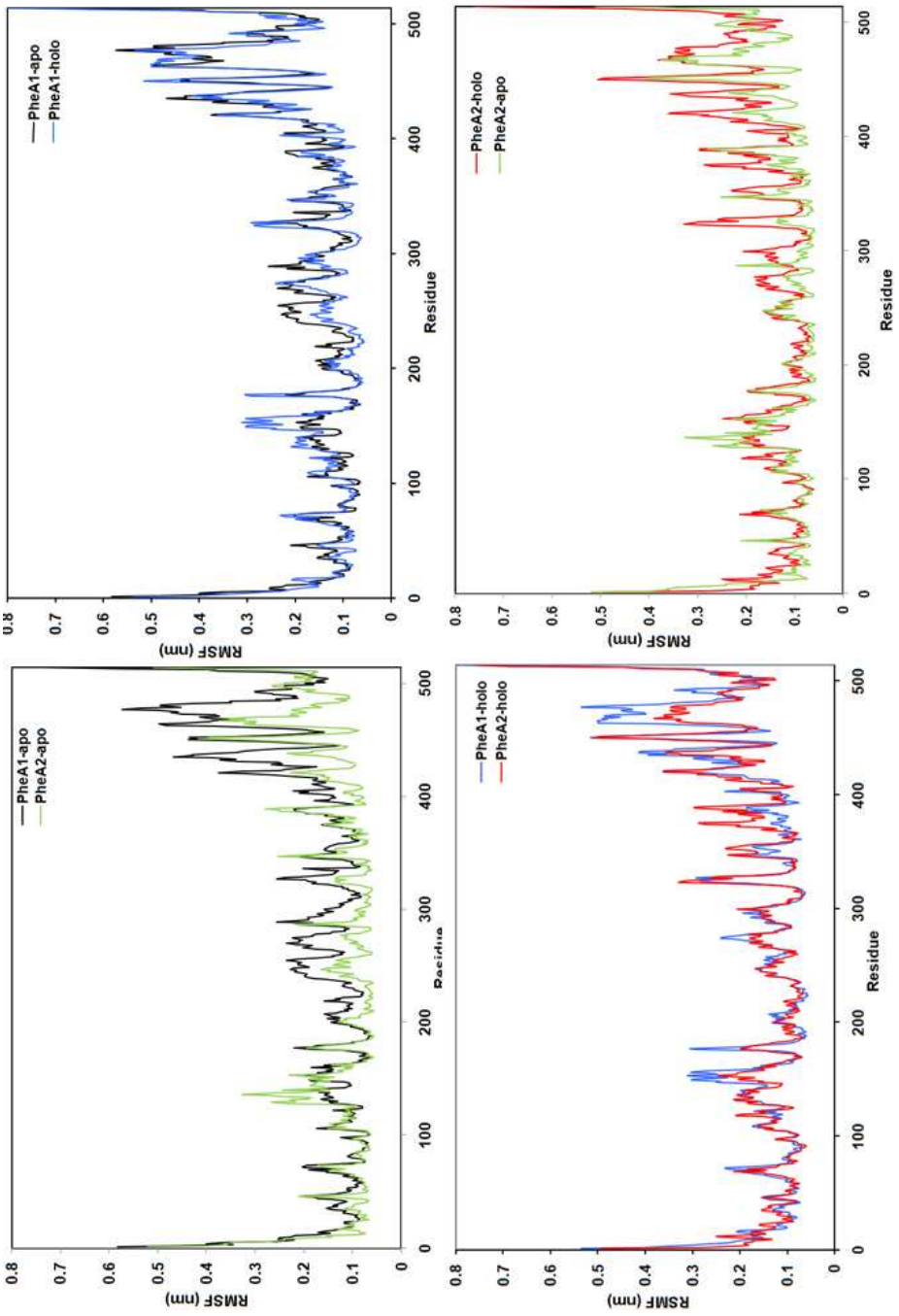


Figure 3.6: **RMSF PheA1-apo, PheA2-apo, PheA1-holo and PheA2-holo simulations.** The time-averaged Cα RMSFs as a function of residue number for the PheA1-apo simulation (black), PheA1-apo simulation (green), PheA1-holo simulation (blue), and PheA2-holo simulation (red) where residue 1 corresponds to pdb residue 16. Comparisons of the Phe1A-apo and PheA2-apo RMSFs (upper left graph), PheA1-holo and PheA2-holo RMSFs (lower left graph), PheA1-apo and PheA1-holo RMSFs (upper right graph) and PheA2-apo and PheA2-holo RMSFs (lower right graph) are shown.

- helix H7, formed by residues 468 to 479, which links β-sheets 2 and 3 of subdomain E of the A_{sub} domain

Residues in the A_{sub} domain of the PheA2-apo simulation are significantly less flexible than those in the PheA1-apo simulation and either of the holo simulations. This, coupled with the observation from the RMSD of the Cα atoms, indicates less movement of the domains of PheA relative to one another in the PheA2-apo simulation as compared with the other simulations.

A number of other regions of flexibility of PheA were indentified in each simulation and will be discussed in relation to the suggested biological significance of each region. In all simulations the greatest flexibility is apparent in the sharp turns between secondary structural elements and flexible loop regions.

The A3 motif loop exhibits the greatest flexibility in the PheA1-holo simulation. The magnitude of the flexibility of this region is comparable in both PheA apo and PheA2-holo simulations.

Greater flexibility of the residues from the L-Phe substrate binding pocket (219–315, pdb: 235–331) is observed in the PheA1-apo simulation as compared to the PheA2-apo and PheA holo simulations. The flexibility of this region in the PheA1-apo simulation can be attributed to the fact that the structure used to initiate the MD was crystallised with ligands present. The fluctuations of this region in the PheA2-apo simulation are however comparable to those of the PheA holo simulations.

Residues 145–166 (pdb: 161–182) of PheA form a long unstructured region that links strand A4 to strand A5. These were identified from the DSSP analysis as forming a very short α-helix in simulations PheA2-apo, PheA1-holo and PheA2-holo, and intermittently in PheA1-apo. This unstructured region exhibits greater flexibility in the PheA-holo structures than in the PheA-apo structure, and more flexibility in the PheA1-holo structure than in the PheA2-holo structure.

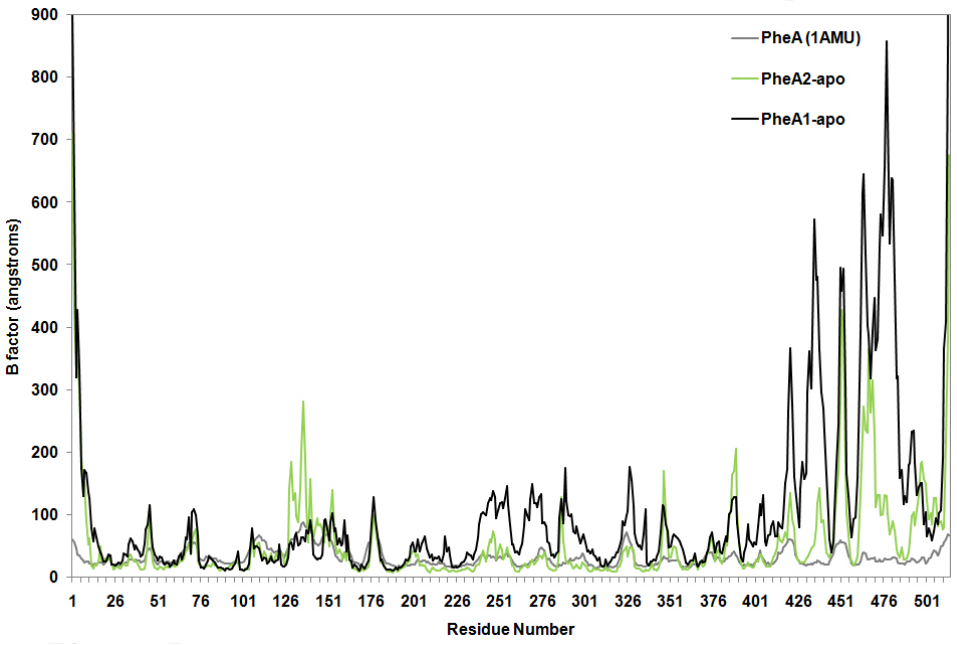


Figure 3.7: **Comparison of PheA1-apo and PheA2-apo RMSF with PheA B-factors.** Normalised time-averaged RMSFs of the Cα atoms of PheA in the PheA1-apo simulation (black) and PheA2-apo simulation (green) as a function of residue number, plotted against the B-factors from the PheA enzyme PDB record (grey).

Comparison with B-Factors

The Cα atom RMSF values from each simulation were normalised to enable comparison with the B-factors from the PheA PDB record. B-factors measure the dynamic disorder caused by the temperature-dependent vibration of the atoms providing information on the relative vibrational motion of different parts of the structure. The comparisons of the B-factors with the normalised Cα atom RMSF values for the apo and holo simulations are shown in figures 3.7 and 3.8 respectively.

Qualitatively, these curves agree for the majority of residues, with the highest fluctuations being seen in the extracellular loops and the lowest fluctuations in the core regions of secondary structure of the A_{core} domain. Two main differences are observed when considering comparison of the curves for the entire protein.

Firstly, the peak values of the RMSFs for the loops of the A_{core} domain are higher in the simulations than in the crystal structure. Secondly, the normalised RMSFs from the simulations are signicantly higher for the core regions of the A_{sub} domain than those derived from the B-values of the PheA crystal structure. This likely reflects constraints on this region of

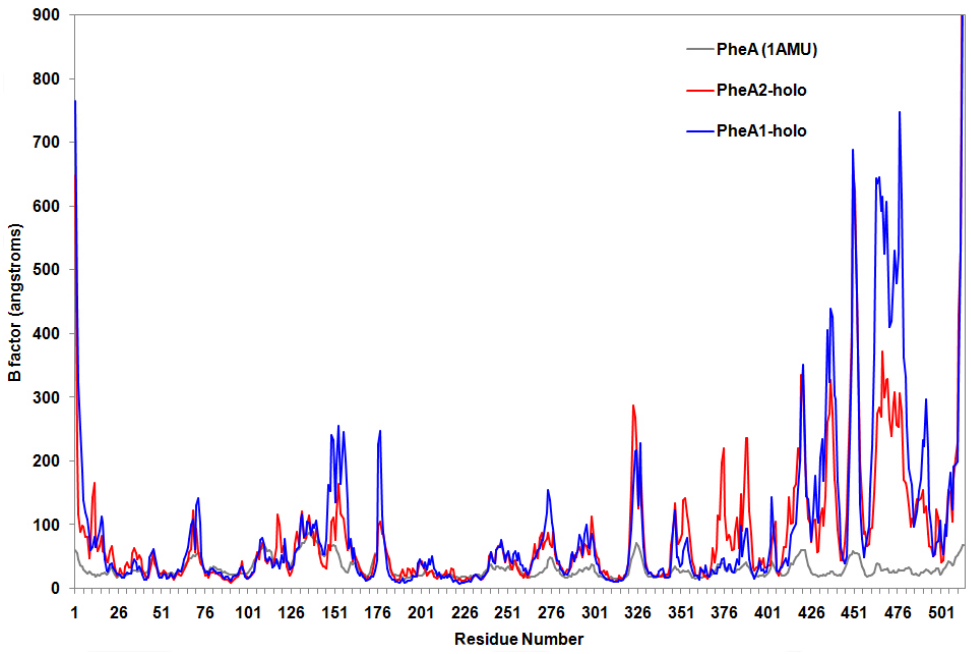


Figure 3.8: **Comparison of PheA1-holo and PheA2-holo RMSF with PheA B-factors.** Normalised time-averaged RMSFs of the Cα atoms of PheA in the PheA1-holo simulation (blue) and PheA2-holo simulation (red) as a function of residue number, plotted against the B-factors from the PheA enzyme PDB record (grey).

the structure present in the crystal which are removed when the protein is simulated in an aqueous environment at a biologically relevant temperature.

3.3.5 Principal Modes of Motion

The principal modes of motion of PheA in each simulation were identified using PCA analysis. The backbone or Cα atoms are commonly used for the fitting in a PCA as these subsets of atoms capture most of the conformational change in the protein. One alternative to assess interdomain motion in the PheA A domain would be to perform a PCA for each system after least squares fitting to the backbone atoms of the larger A_{core}; atoms 23 to 414 (ppdb: 39–430). This was considered, however, by fitting to the A_{core} domain, fluctuations of the A_{sub} domain relative to this domain would be exaggerated and the analysis restricted to only identifying relative A_{core}/A_{sub} domain motion; this option was therefore not selected.

Figure 3.9 describes the size of each of the ten first eigenvectors (index) sorted by size, with the first having the largest eigenvalue. As expected, only the first two eigenvectors

Index	PheA1-apo Eigenvalue Cumulative (nm^2)	(%)	PheA2-apo Eigenvalue Cumulative (nm^2)	(%)	PheA1-holo Eigenvalue Cumulative (nm^2)	(%)	PheA2-holo Eigenvalue Cumulative (nm^2)	(%)
1	28.94	48.70	11.98	39.04	25.91	46.66	18.41	38.29
2	10.17	65.81	2.46	47.06	8.13	61.31	10.64	60.42
3	2.93	70.74	1.83	53.00	3.17	67.01	2.54	65.69
4	1.98	74.06	1.45	57.73	1.91	70.45	1.83	69.51
5	1.78	77.06	1.04	61.19	1.56	73.26	1.64	72.92
6	0.91	78.60	0.72	63.48	1.32	75.64	1.01	75.02
7	0.76	79.89	0.62	65.50	0.79	77.06	0.79	76.65
8	0.67	81.02	0.61	67.49	0.76	78.42	0.64	77.98
9	0.61	82.05	0.54	69.26	0.64	79.57	0.56	79.14
10	0.53	82.94	0.42	70.61	0.53	80.53	0.51	80.20

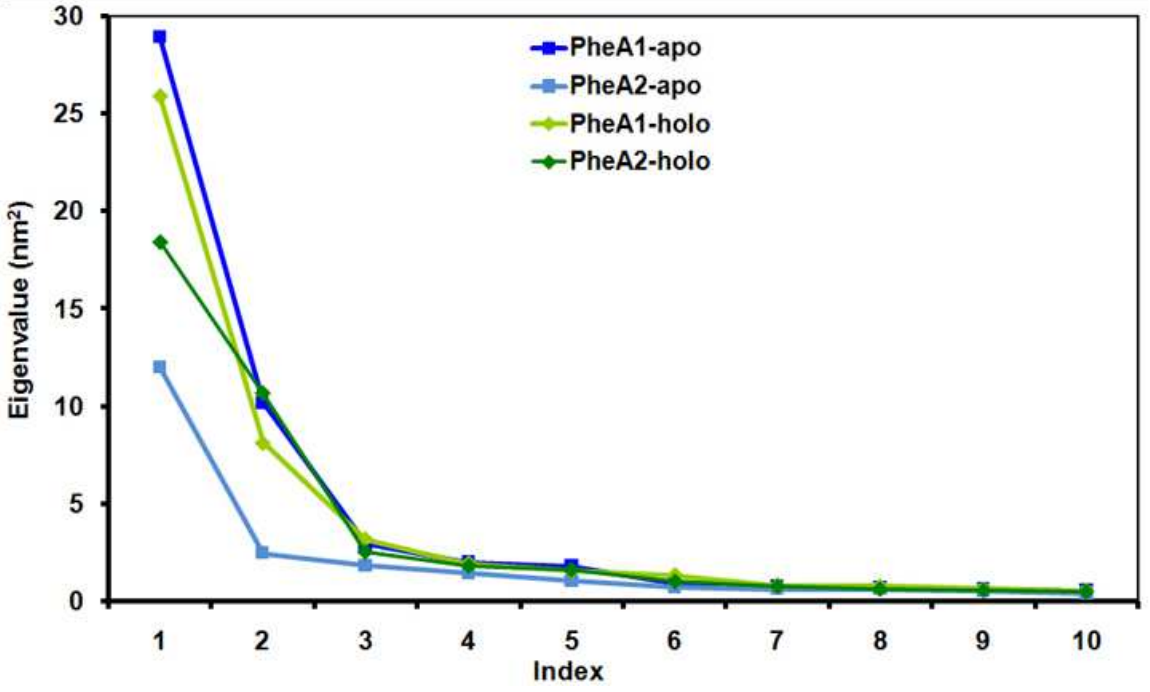


Figure 3.9: **PCA analysis of the PheA apo and holo simulations**The eigenvectors (index) and eigenvalues of the PheA1-apo (blue), PheA2-apo (light blue), PheA1-holo (light green), and PheA2-holo (green) simulations.

Simulations	PheA1-apo	PheA2-apo	PheA1-holo	PheA2-holo
PheA1-apo	0.59	0.55	0.62	0.65
PheA2-apo		0.56	0.55	0.55
PheA-holo			0.62	0.59
PheA2-holo				0.67

Table 3.5: RMSIP between the first ten eigenvectors for the PheA-apo and holo simulations.

have large eigenvalues.

It is interesting to note that the magnitude of the first ten eigenvectors for the first set of simulation (PheA1-apo and PheA1-holo) is very similar, with greater dissimilarity between the PheA1-holo and PheA2-holo eigenvectors distribution. The first eigenvector of the PheA2-holo simulation is significantly lower than that of the PheA1-holo simulation, conversely the second eigenvector from the PheA2-holo simulation is higher than that of the PheA-holo simulation. Cumulatively 46.6% and 61.6%, and 38.42% and 60.42% of the total motion in the PheA1-holo and PheA2-holo simulation is described by the first and second eigenvectors respectively.

The magnitude of the eigenvectors for the PheA2-apo simulations is quite different to the PheA1-apo simulation, with significantly lower values observed for the first two eigenvectors. The RMSD and RMSF analysis of the trajectory of this simulation indicates less motion between the domains in PheA in this simulation and greater stability in the original positioning of the domains in the PheA2-apo simulation as compared to the starting structure.

The similarity in the motions between the four proteins can be determined by calculating the root mean square inner product (RMSIP) of the eigenvectors. The RMSIP measures the overlap of the motions for each protein in the subspace spanned by the respective eigenvectors. The RMSIP was calculated for the first ten eigenvectors for PheA between each of the four simulations, shown in table 3.5. For reference, the RMSIP was also calculated for PheA in each individual simulation by splitting the trajectories in half and comparing the first two eigenvectors from each half of the trajectory (these values are shown on the diagonal in the table). The values of RMSIP show that essential subspaces of the protein are overlapped.

To analyse the nature of the collective motions of PheA in each system the trajectories from each principal component analysis were projected onto the respective first three eigenvectors to reveal the sampling along these vectors. The extreme projections of the trajectories along the first three eigenvectors were obtained. These structures were processed using the DynDom server. The DynDom program and visual inspection of the conformations which correspond to the extremes of the projection of the eigenvectors onto the trajectory were used to identify the nature of the motion corresponding to the principal eigenvectors. The trend of each motion will be described and a comparative summary provided at the end of the section.

3.3.6 Principal Modes of Motion - Holo Simulations

PheA1-holo EV1

The modes of motion of the first three eigenvectors of PheA from the PheA1-holo simulation are shown in figure 3.10.

The extreme conformations of the motion described by eigenvector one in the PheA1-holo simulation are evident at 0.049 and 7.005 ns. This eigenvector largely describes motion between the A_{core} and subdomain D of the A_{sub} domain, and helix H6 and subdomain E of the A_{sub} domain. Nine residues from the A_{sub} domain do not move in a concerted fashion with the rest of the domain. These residues are 417–425 (pdb: 432–441), A8 motif residues which form subdomain D, and 510–511 (pdb: 526–527). It is perhaps unsurprising that residues 510 and 511 are considered part of the A_{core} domain; this region of the structure is effectively anchored in place directly above the cleft between the two domains through the interaction of Lys 501 with the L-Phe and AMP ligands. Two residues (175–176, pdb: 191-192), from the A3 motif loop, from in the A_{core} domain move in concert with the A_{sub} domain. This region was identified as being highly flexible in this simulation during the RMSF analysis.

Bending residues (suggested hinge regions) for this motion are residues 417–426 (which

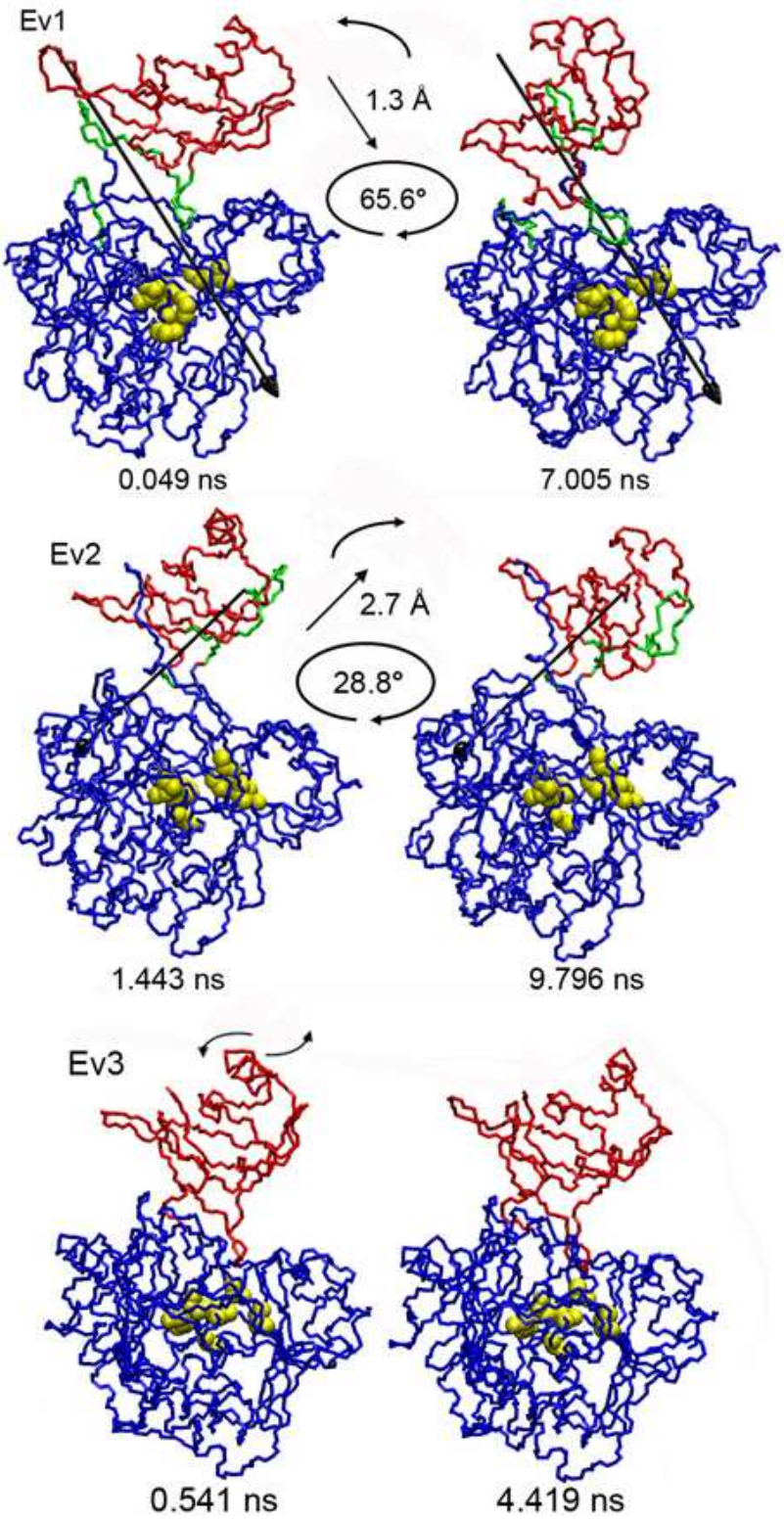


Figure 3.10: **Domain motion in PheA1-holo.** Interdomain motion in the PheA1-holo simulation; corresponding to the first three eigenvectors and the first two of which were identified by DynDom. Domain 1 (static) is shown in blue, domain 2 (moving) in red, the hinge regions in green and the phenylalanine binding pocket residues in yellow using VDW representation.

follow the highly conserved L-Asp residue (414, pdb: 430) that form part of motif A8, residues 174–179 from the A3 motif loop and residues 496–502, which contain the A10 motif residues. The motion described by this eigenvector is a 65 ° rotation and translation of 1.3 $\mathring{A}$ of the A_{sub} domain (excluding subdomain D) about the axis defined by the hinge residues, resulting in the domain twisting towards the centre of the active site cleft and the A3 motif loop located on the left side of the protein. Moving in concert with the A_{sub} domain, the A3 motif loop twists increasing exposure of the binding pocket of the ligands.

PheA1-holo EV2

The second eigenvector from the PheA1-holo simulation describes the tilting of the A_{sub} domain towards the right of the protein, away from the A3 motif loop. The extreme projections of the motion described by eigenvector 2 are evident at 1.443 and 9.796 ns. As for eigenvector 1 the static domain is largely comprised of the residues from the A_{core} domain, with a small number of residues from the A_{sub} domain; residues 441 to 443 (pdb: 457 to 459), and 494 to 511 (pdb: 514 to 527). The hinge regions for this motion are residues 413–414 (motif A8 residues), which include the highly conserved L-Asp residue (414, pdb 430), residues 432 to 441, 443 to 444, 493 and 494. This motion can be described as a 2.7 $\mathring{A}$ translation and 29° rotation of the A_{sub} domain about the axis.

PheA1-holo EV3

No domain motion was identified using DynDom for eigenvector three of the PheA1-holo simulation. Visual comparison of the extreme conformations of PheA that correspond to this eigenvector suggest the largest motion described by this eigenvector is the twisting of helix E1 of structural subdomain E of the A_{sub} domain.

PheA2-holo EV1

The modes of motion of the first three eigenvectors of PheA2-holo are shown in figure 3.11.

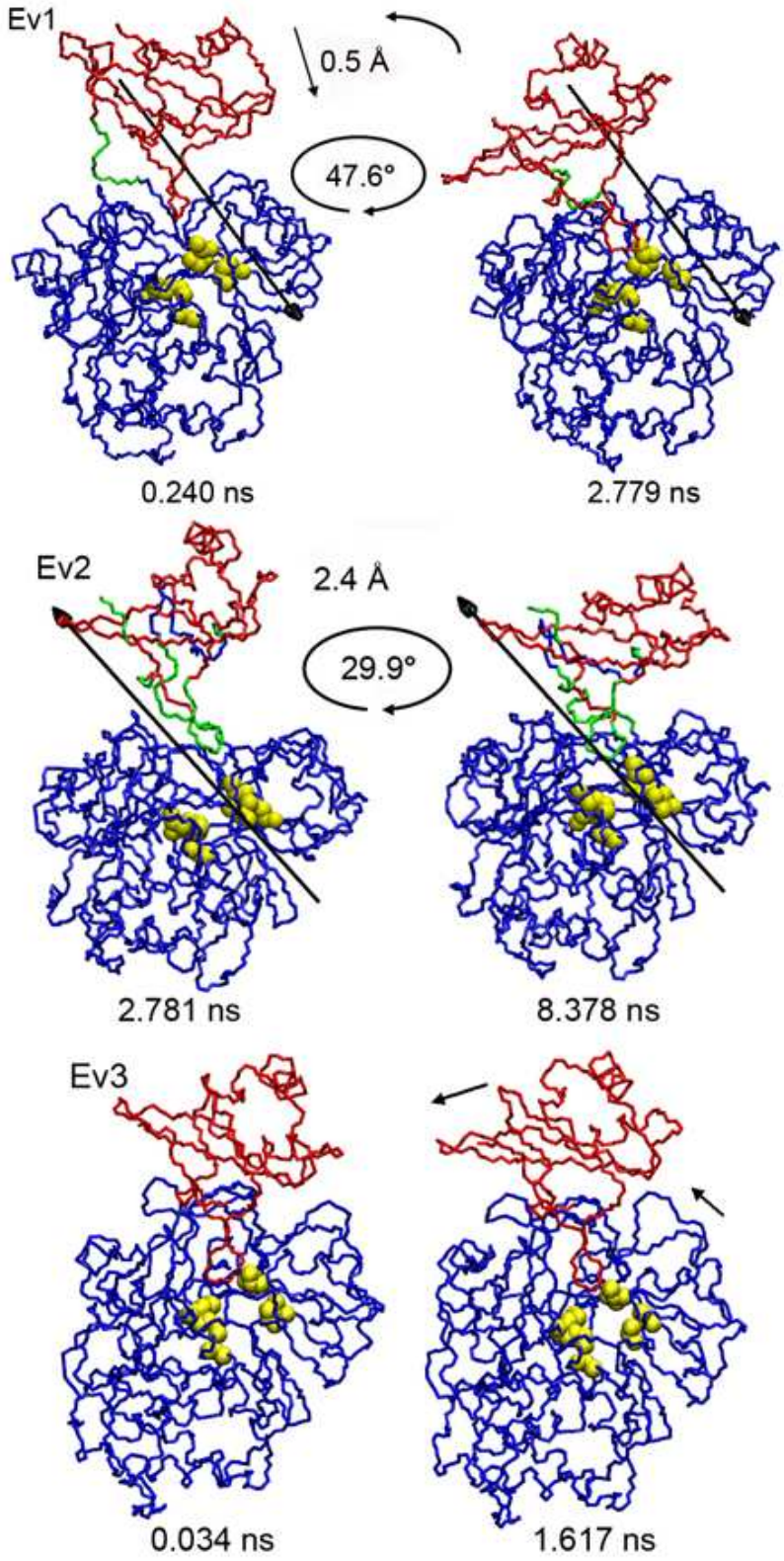


Figure 3.11: **Domain motion in PheA2-holo.** Interdomain motion in the PheA2-holo simulation; corresponding to the first three eigenvectors and identified by DynDom. Domain 1 (static) is shown in blue, domain 2 (moving) in red, the hinge regions in green and the phenylalanine binding pocket residues in yellow using VDW representation.

The extreme conformations of eigenvector one for the PheA2-holo simulation are evident at 0.240 and 2.779 ns. This eigenvector describes the motion of the A_{sub} domain (residues 415–512, pdb: 431–530) relative to the A_{core} domain (residues 14–414, pdb: 30–430) about the hinge region (residues 411–416 from motif A8 and which include the Arg 412 and Asp 414 residues) which links the two domains. The direction of motion is similar to that described by eigenvector 1 in the PheA1-holo simulation; the A_{sub} domain tilts towards the A3 motif loop on the left of the PheA and twists in a clockwise direction about the hinge axis. This motion is of magnitude -0.5 $\AA$ translation and 48° rotation about the hinge axis.

PheA2-holo EV2

The conformations corresponding to the extremes of the motion described by eigenvector 2 are observed at 2.781 and 8.378 ns. This eigenvector describes the motion of subdomain E and part of helix H6 of the A_{sub} domain relative to the A_{core} domain and subdomain D and part of helix H6 of the A_{sub} domain in a clockwise direction and tilting slightly towards the right of the protein, however the overall motion brings this moving region of the A_{sub} domain closer to the A_{core} domain, almost in a lid closing motion. Suggested hinge residues for this motion include 410–417 (A8 motif), 429–430, 495–501 (including A10 motif), 505–506 and 509–511.

PheA2-holo EV3

No domain motion was identified from eigenvector 3 of the PheA2-holo simulation. Overlay of the two structures corresponding to the extremes of this eigenvector suggest flexibility in three loops at the interface of the domains; one from the A_{core} domain and two from the A_{sub} domain.

3.3.7 Principal Modes of Motion - Apo Simulations

The apo state structure for these simulations was taken by removing the ligands from the holo state crystal structure. The modes of motion and behaviour observed in these simulations therefore may not be reflective of the behaviour of the full apo state protein. In the PheA structure the long loop (K loop) from the A_{sub} domain, which contains the invariant lysine residue, is anchored in the binding pocket of the protein. While the conformation of PheA is very similar to that of Dhbe, which was determined without ligands, comparison of PheA with the A domain from SrfA-C (crystallised with L-Leu), see figure 1.11 in Chapter 1, reveals a more open structure where the A10 motif K loop is lifted from the binding pocket. The motion of PheA in the apo simulations presented in this chapter can, however, provide a comparison of the effect of the ligands on the dynamical behaviour of PheA.

PheA1-apo motion

Motion between domains was identified by DynDom from the extreme conformations of the first three eigenvectors of the PheA1-apo simulation. These motions are shown in figure 3.12.

PheA1-apo EV1

The extreme conformations described by the first eigenvector of PheA1-apo were identified at 0.718 and 11.036 ns. This eigenvector describes motion of the A_{sub} domain relative to the A_{core} and six residues from the A10 motif K loop. Motif A8 (413–414) and A10 residues (496–500 and 502–506) were identified as hinge residues for this motion. This eigenvector describes a translation of -1.0$\mathring{A}$ and rotation of 42° about the axis defined by these hinge residues which correlates to the A_{sub} domain twisting clockwise, slightly backwards (opening of the binding cleft) and tilting to the right, towards the A_{core} domain and away from the A3 motif loop.

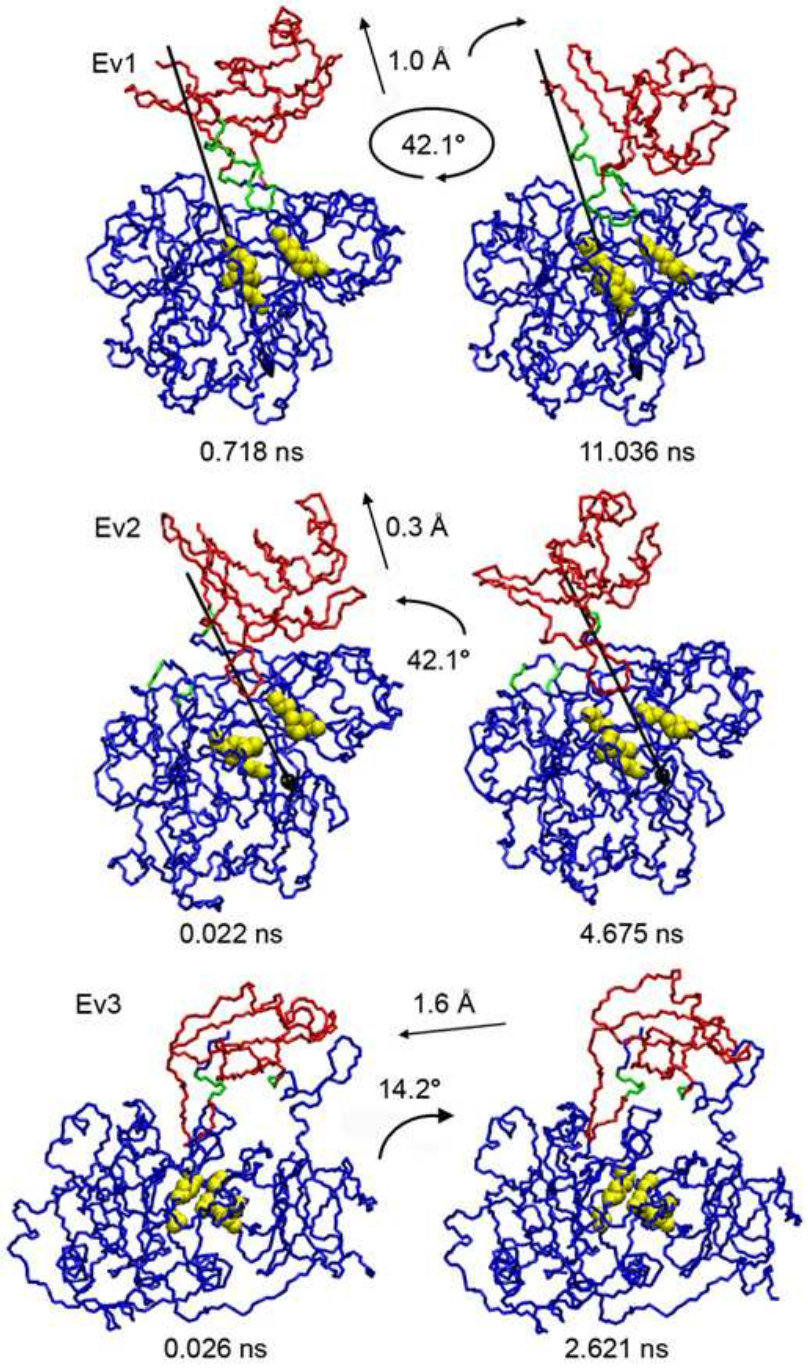


Figure 3.12: **Domain motion in PheA1-apo.** Interdomain motion in the PheA1-apo simulation; corresponding to the first three eigenvectors and as identified by DynDom. Domain 1 (static) is shown in blue, domain 2 (moving) in red, the hinge regions in green and the phenylalanine binding pocket residues in yellow using VDW representation.

PheA1-apo EV2

The second eigenvector of the PheA1-apo simulation describes motion between the A_{core} and A_{sub} domain. The conformations corresponding to the extremes of this motion were identified at 0.022 and 4.675 ns. Domain 1 comprises residues 11–175 and 180–415, domain 2 comprises residues 176–179 and 416–511 and the hinge residues are defined as 175–176, 179–180 (from the A3 motif), and 415–416 (from the A8 motif). Residues 177–178 of the A3 motif loop move in concert with the A_{sub} domain. This eigenvector describes a 42° rotation and translation of -0.3 $\AA$ of the A_{sub} domain about the axis. This motion is the A_{sub} domain twisting clockwise and tipping towards the left of the A_{core} domain (towards the A3 motif loop). As this happens the A_{sub} domain tilts backwards slightly; possibly to accommodate this rotation and tipping movement.

PheA1-apo EV3

The third eigenvector of PheA1-apo describes the motion between the A_{core} domain and subdomain D and part of helix H6 from the A_{sub} domain, and part og helix H6 and subdomain E of the A_{sub} domain. The conformations corresponding to the extremes of this motion were identified at 0.026 and 2.621 ns. Domain 1 comprises residues 13–430 and 504–509, domain 2 is comprised of residues 431–503, and the hinge residues are 430–431 and 503–506. This motion can be described as 14° rotation and translation of 1.6 $\AA$ of sudomain E of the A_{sub} domain about the axis. This eigenvector describes lifting of the A_{sub} subdomain E from the A_{sub} subdomain D and the A_{core} domain which remain static. As this happens, subdomain E tilts back away from the binding cleft lifting the A10 motif loop from the binding pocket, slightly exposing the binding cleft.

PheA2-apo motion

DynDom identified domain motion in the first three eigenvectors from the PheA2-apo simulation. The modes of motion are shown in figure 3.13.

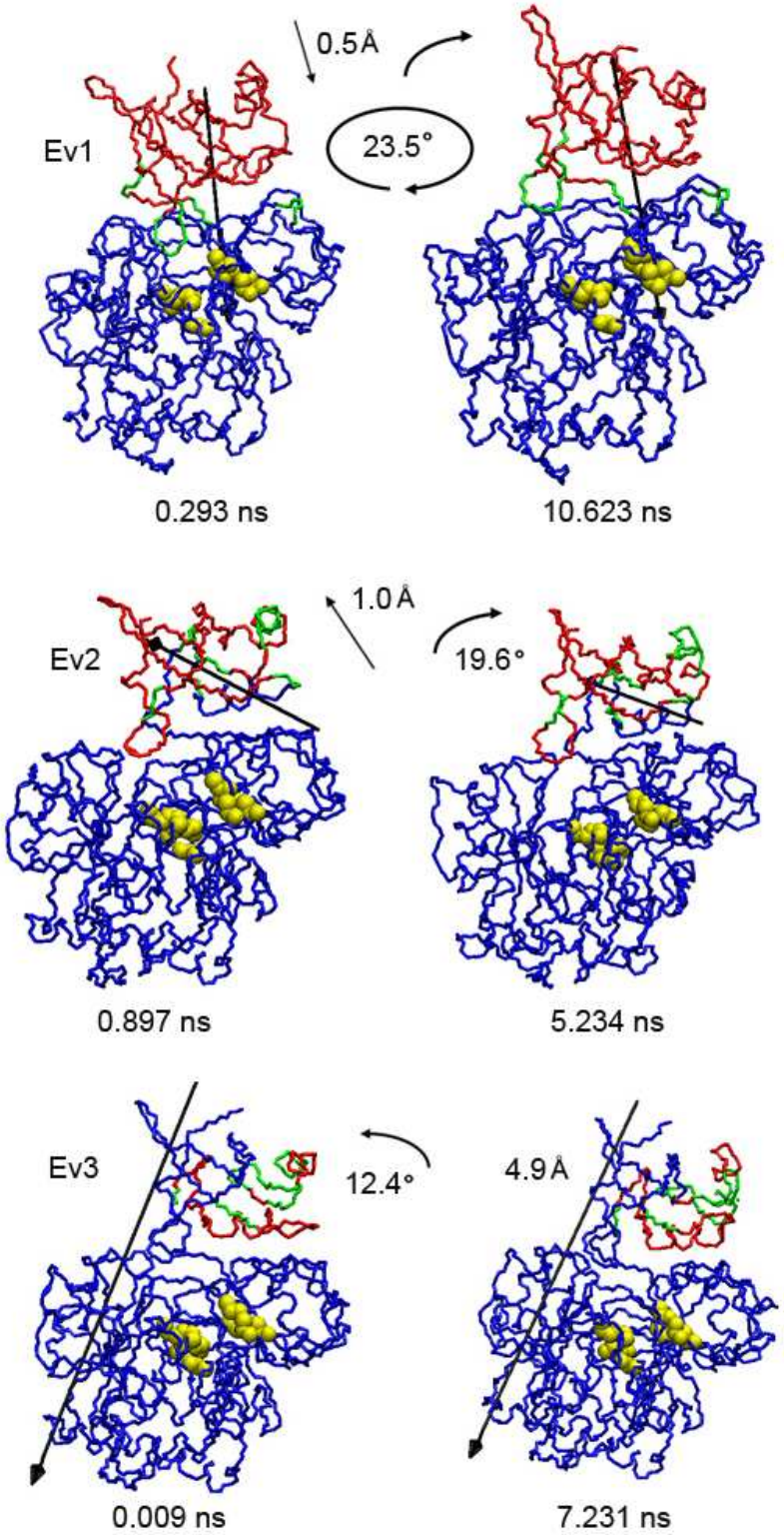


Figure 3.13: **Domain motion in PheA2-apo.** Interdomain motion in the PheA2-apo simulation; corresponding to the first three eigenvectors and as identified by DynDom. Domain 1 (static) is shown in blue, domain 2 (moving) in red, the hinge regions in green and the phenylalanine binding pocket residues in yellow using VDW representation.

PheA2-apo EV1

The first eigenvector of PheA2-apo describes motion between the A_{core} and A_{sub} domain. The extreme conformations of this motion were identified at 0.293 and 10.623 ns. Domain 1 comprises residues 3–291, 293–409, 415, 498–499, and 501, and domain 2 residues 292, 410–414, 416 - 497, 500 and 502 - 512. Residues 291 - 293, 409 - 410, 414 - 416 (from motif A8), and 497 - 502 (from motif A10) were identified as forming the hinge between the moving domains. This eigenvector describes a 24° rotation and translation of 0.5 $\AA$ of the A_{sub} domain about the hinge axis. As was seen in the DynDom analysis of the extreme conformations of the first eigenvector of the PheA1-apo simulation, a number of residues from the A 10 motif K loop do not move with the A_{sub} domain. In addition, three residues (291 - 293) of the A_{core} domain move in concert with the A_{sub} domain. This eigenvector describes a combination of the A_{sub} domain twisting clockwise, and tilting to the right, towards the A_{core} domain, and away from the A3 motif loop. Figure 3.13 illustrates this motion and also shows a slight lifting of the A10 motif loop from the binding cleft.

PheA2-apo EV2

The second eigenvector of the PheA1-apo simulation describes motion between the A_{core}, structural subdomain D and helix H6 of the A_{sub} domain, and structural subdomain E of the A_{sub} domain. Domain 1 comprises residues 6–419, 421–423, 425–437, 460–460, 466–472, and 496–503. Domain 2 comprises residues 420, 424, 438–459, 461–465, 473–495, and 504–509. The hinge residues for this motion are 419–421, 423–425, 437–438, 459–461, 465–474, 495–496, and 503–504. This eigenvector describes a 20° rotation and translation of 1.0 $\AA$ of domain 2 about the axis, which equates to a tipping of structural subdomain E of A_{sub} domain towards the right side of PheA, and away from the A3 motif loop.

PheA2-apo EV3

No concerted domain motion was identified from the DynDom analyses of the extreme conformations from eigenvector 3.

3.3.8 Principal Modes of Motion - Comparison

While the largest motion of domain in each holo simulation occurs in a similar direction, the region of the A domain that is moving is different.

The first eigenvector of the PheA1-holo simulation describes the rotation of the subdomain E and helix H6 of the A_{sub} domain in a clockwise direction and tilting towards the left side of PheA about an axis defined by residues from the A3, A8 (417–425) and A10 motifs.

In comparison the principal motion observed in the PheA2-holo simulation describes the rotation of the entire A_{sub} domain in a clockwise direction and tilting towards the left side of PheA about an axis defined by the A8 motif residues. The hinge residues for this motion are 411–416 which include Arg 412 and Asp 414. An arginine residue equivalent to Arg 412 in PheA (pdb: 428) was identified by Dieckmann and co-workers from limited proteolysis of TycA as being a site of intrinsic flexibility, which decreased in the presence of the ligands[75,113]. Comparison of first-half and second-half reaction structures of members of the adenylate-forming superfamily identified the conserved aspartic acid residue (PheA 414, pdb: 430), the first Asp residue of the A8 motif GRx**D**xQVKIRGxRIELGEIE, as the hinge about which domain alternation occurs[52,53].

The split of domains between which the second principal motion occurs in the holo simulations is reversed and the direction of motion is different.

The second eigenvector from the PheA1-holo domain describes motion of the A_{sub} domain tilting towards the right side of PheA away from the A3 motif loop residues, about an axis defined by the A8 (413–414) motif residues and residues 432 to 441, 443 to 444, 493 and 494. The right and left sides of PheA are illustrated in figure 3.14. In this simulation the

A3 motif loop residues are more flexible than in the PheA2-holo simulation.

The second eigenvector from the PheA2-holo domain describes the motion of subdomain E and part of helix H6 from the A_{sub} domain in a clockwise direction and tilting slightly towards the right side of PheA, however, the overall motion brings this moving region of the A_{sub} domain closer to the A_{core} domain. This motion occurs about an axis defined by the A8 motif (410–417) residues and residues 429–430, 495501 (residues from the A10 motif), 505–506 and 509–511. The A3 motif residues in this simulation are not as flexible as observed in the PheA1-holo simulation.

In each of the principal modes of motion from the PheA-holo simulations residues from the A8 motif act as a hinge. The A10 motif residues are also indicated as hinge residues in some of the motions; this is perhaps unsurprising given these residues interact with the L-Phe substrate and AMP ligands, anchoring this region of the structure in the active site.

The principal motion in each apo simulation occurred between the A_{sub} and A_{core} domain.

In the PheA1-apo simulation this motion described the tipping of the A_{sub} domain twisting clockwise, slightly backwards (opening of the binding cleft) and tilting to the right, towards the A_{core} domain and away from the A3 motif loop. In this motion residues from the A8 motif A8 (413–414) and A10 motif (496–500 and 502–506) were identified as hinge residues.

The interdomain motion described by the first eigenvector of each apo simulation is qualitatively similar, with the A_{sub} domain tipping away from the A3 motif loop. This motion in the PheA2-apo simulation is however accompanied by a slight lifting of the A10 motif K loop and in the PheA1-apo simulation by the backwards tilting of the A_{sub} domain.

The second eigenvector of the PheA1-apo simulation describes motion between the A_{core} and A_{sub} domain with the A_{sub} domain twisting clockwise and tipping towards the left of the A_{core} domain (towards the A3 motif loop). As this happens the A_{sub} domain tilts backwards slightly; possibly to accommodate this rotation and tipping movement. The division of domains in the PheA2-apo simulation is different and similar to that observed in the PheA2-holo simulation with the motion occurring between the A_{core}, structural subdo-

	PheA1-apo	PheA2-apo	PheA1-holo	PheA2-holo	
Starting	761	746	754	762	
Whole simulation (SD)	741.3 (20.0)	744.6 (17.2)	724.6 (18.8)	747.5 (21.1)	
1st ns (SD)		718.7 (15.9)	731.1 (15.8)	722.0 (17.0)	727.1 (16.7)
Final ns (SD)		761.8 (15.6)	756.7 (15.6)	726.6 (14.2)	755.5 (16.5)

Table 3.6: Average number of intramolecular hydrogen bonds (P-P H bonds) for the apo and holo PheA simulations.

main D and helix H6 of the A_{sub} domain, and structural subdomain E of the A_{sub} domain. This motion equates to a tipping of structural subdomain E of A_{sub} domain towards the right side of PheA, and away from the A3 motif loop.

In each motion observed in the apo simulations residues from the A8 motif, as well as residues from other regions including the A10 and A3 motifs, act as hinge residues.

3.3.9 Intramolecular Hydrogen Bonding

The average number of intramoleclar (protein-protein) hydrogen bonds was obtained for PheA for the whole simulation, frist nanosecond and last nanosecond for each simulation, see table 3.3.9.

The average number of intramolecular hydrogen bonds in the PheA is largely consistent across the simulations with the exception of the PheA1-holo simulation where there are on average 20 less hydrogen bonds between PheA are present per picosecond than in the other simulations. Plotting the average number of hydrogen bonds against time for each simulation shows that the average number of hydrogen bonds gradually increases throughout the PheA1-apo, PheA2-apo and PheA2-holo simulations. However in the PheA1-holo simulation the average number of hydrogen bonds decreases during the first five nanoseconds, increases during the sixth and seventh nanoseconds to a peak at an average 744 bonds and then decreases over the remainder of the simulation. The greatest average number of intramoelcular hydrogen bonds observed in the PheA1-holo simulation correlates with the time that the extreme of motion described by the first eigenvector is observed.

3.3.10 Interdomain Hydrogen Bonding

Given the observed motion of the domains relative to one another in each simulation an analysis of the interdomain hydrogen bonding was performed, using distance and angle criteria of 3.6 $\AA$ and 60°. For this analysis the A_{core} domain was defined as comprising residues 1 to 414 (pdb: 17 to 430) and the A_{sub} domain the remaining residues from the protein.

Analysis of interdomain hydrogen bonds with reference to the PheA structure revealed interactions between the two domains clustered around four distinct regions.

- the interdomain hinge;
- the A_{sub} domain motif A10 K loop;
- the A3 motif loop and residues on the left side of PheA, and
- the right side of PheA.

The orientation of PheA used to define the left and right hand sides of PheA is shown in figure 3.14. Hydrogen bonds were defined as being in one of these clusters if a hydrogen bonding interaction was formed between any residue within the region.

The average number of hydrogen bonds per nanosecond was calculated and will be used as a measure of the hydrogen bonding strength between particular regions of the protein. The average hydrogen bonding per cluster per nanosecond versus time are shown figures 3.15. This analysis shows that the location of the interdomain hydrogen bonds formed between the two domains broadly correlates with the motion described by the principal eigenvectors.

3.3.11 Interdomain Hydrogen Bonding - Holo Simulations

In the holo simulations the majority of interdomain hydrogen bonding is observed between between residues from the left side of PheA (including the A3 motif loop), and residues in the hinge region, with fewer hydrogen bonds observed forming between residues the

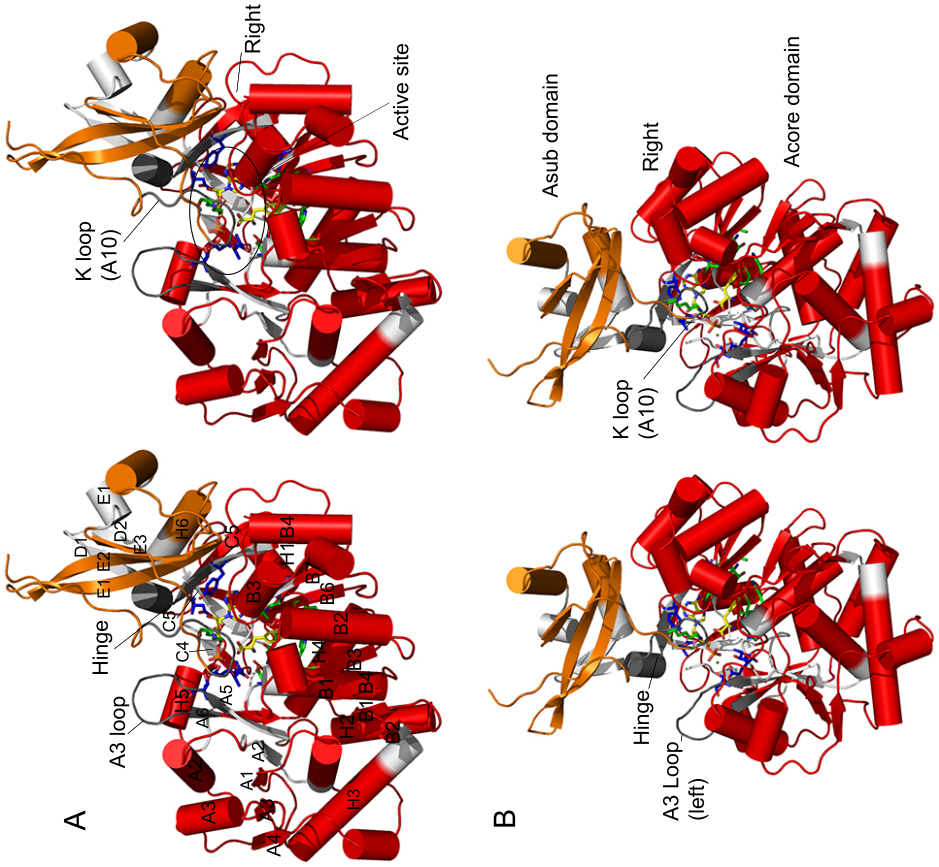


Figure 3.14: **PheA interdomain hydrogen bond interaction groupings.** A) The Acore domain is shown in red, A_{sub} domain in orange, conserved motifs are shown in white, AMP and Phe in yellow and Mg in lime. PheA annotated with the secondary structural elements as defined in *Conti et al*[62]. B) The 4 regions of interdomain hydrogen bonding are annotated; A3 loop (left), hinge, K loop (A10) and right.

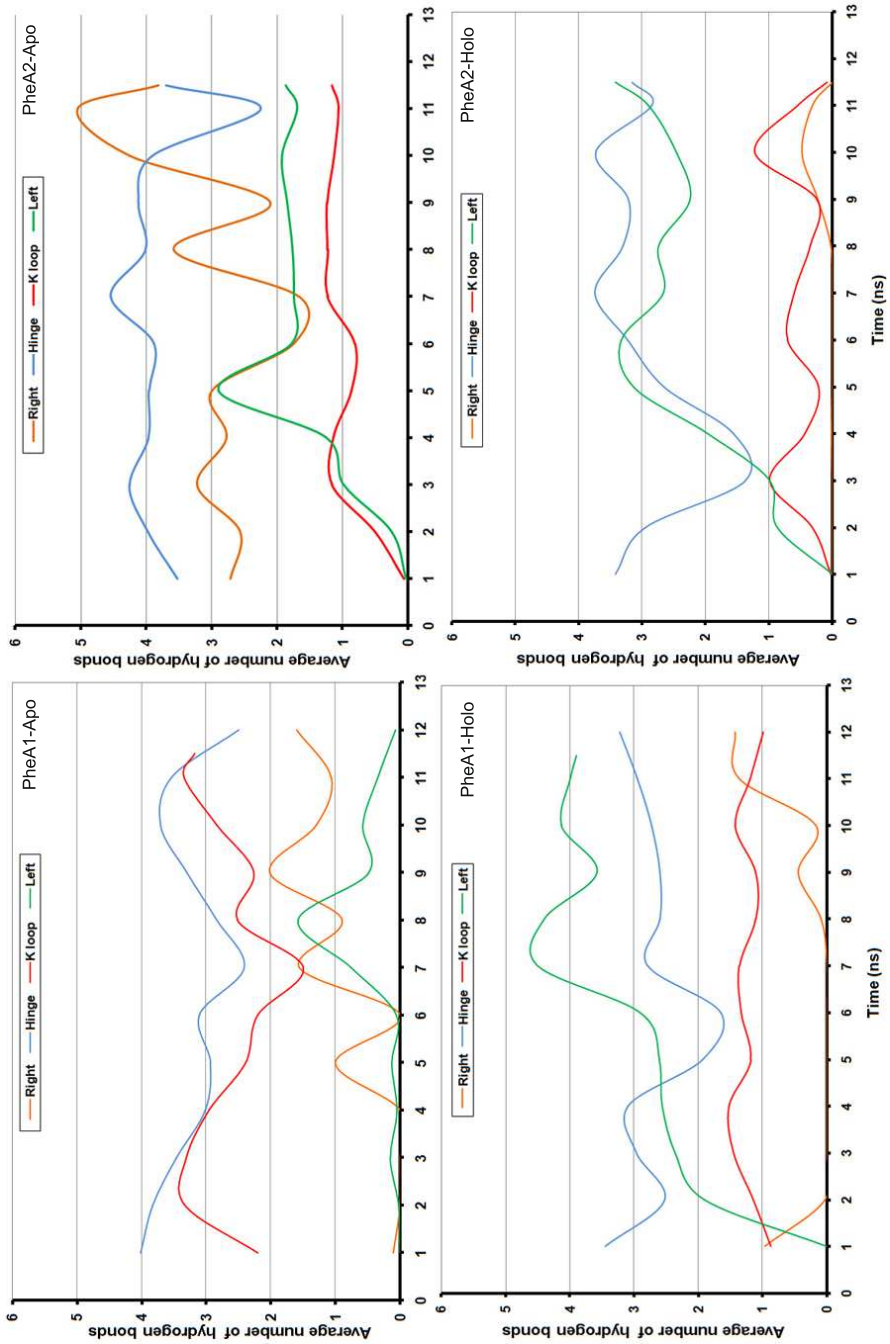


Figure 3.15: **Interdomain hydrogen bonding in the PheA apo and PheA holo simulations.** Graph to show how the average number of interdomain hydrogen bonds between defined regions of the PheA structure fluctuate on a nanosecond timescale in the PheA1-apo (upper left), PheA2-apo (upper right), PheA1-holo (lower left), and PheA2-holo (lower right) simulation. Hydrogen bonds formed at the hinge / interdomain linker region are shown in blue, those formed on the left side of PheA in green, those on the right in orange, and those formed between a residue of the K loop and the PheA A_{sub} domain in red.

right side of PheA. The average number of hydrogen bonds formed on the left side of PheA increases during both simulations; the average number is slightly higher in the PheA1-holo simulation than the PheA2-holo simulation.

Evolution of the average number of hydrogen bonds throughout the holo simulations correlates with the times that the extremes of motion described in the first two principal modes of motion are observed.

In PheA1-holo simulation interdomain hydrogen bonding on the left side of the PheA is strongest ($\sim$4.5) during the seventh nanosecond which is the time the extreme motion described by the first eigenvector occurs. The extreme of the motion of PheA2-holo described by eigenvector 1 is observed at 2.8 ns. At this time the strength of the hydrogen bonding interactions on the left side of PheA begins to increase and is strongest ($\sim$3.5) during the fifth nanosecond. During the eighth nanosecond, when the extreme motion described by the second eigenvector is observed, the hydrogen bonding on the left side of PheA decreases.

Hydrogen bonding interactions on the left side of PheA common to both simulations are formed between residues from the A3 motif loop (Ser 175, Gly 176, and Thr 178) and Thr 497, Arg 504, and Lys 505, from the A_{sub} domain.

3.3.12 Interdomain Hydrogen Bonding - Apo Simulations

Greater motion between the domains is observed in the PheA1-apo simulation than the PheA2-apo simulation and the hydrogen bonding interactions between residues in the hinge region, as compared to the PheA2-holo simulation, correlates with this. Some hydrogen bonding interactions are observed between the residues located on the right side of PheA in the apo simulations, this is consistent with the principal mode of motion for these simulations which describes a tilting of the A_{sub} domain towards the right side of PheA, away from the A3 motif loop.

The fluctuations of strength of interdomain hydrogen bonding at the hinge region over the course of the simulation is very similar to the fluctuations in strength of interdomain hy-

drogen bonding involving at least one residue from the K loop. During the PheA1-apo simulation the strength of the hydrogen bonding between the K loop and A_{core} domain fluctuates between 1.5 and 3.2. The fewest hydrogen bonding interactions are formed at the hinge region and the A10 motif K loop during the seventh ns (6–7 ns). Very few or no hydrogen bonding interactions are present between the domains on the left side of PheA during the first 6 ns. During the eighth ns the strength of interdomain hydrogen bonding on the left side of PheA peaks at 1.7. Very few hydrogen bonding interactions are formed between the domains on the right side of PheA during the first 5 ns. From 6 ns onwards the interdomain hydrogen bonding in this region increases.

The trend of hydrogen bonding interactions at the hinge region during the PheA2-apo simulation is comparable to that observed in the PheA1-apo simulation, although on average stronger interactions are formed at this region in the PheA2-apo simulation. Weaker hydrogen bonding interactions are formed between the A10 motif K loop residues and the A_{sub} domain in the PheA2-apo simulation, however stronger hydrogen bonding is observed are on both the right and left sides of PheA in the PheA2-apo simulation. While the average strength of hydrogen bonding formed between the right and left sides of PheA differs in magnitude between the holo simulations, the overall trend of the hydrogen bonding between this regions throughout the simulation is comparable. Residues on the right side of PheA involved in hydrogen bonding between the domains that are common to both simulations are His 270 from the A_{core} domain, and Ser 440 and Glu 441, from the A_{sub} domain.

3.3.13 Ligand Binding

PheA was crystallised with the products of the first half reaction; L-Phe and AMP. The structure of PheA, and the L-Phe and AMP binding pockets can be seen in figure 3.16.

The L-Phe substrate binding pocket comprises ten residues (pdb numbering in parentheses) Asp 219 (235), Ala 220 (236), Trp 223 (239), Thr 262 (278), Ile 283 (299), Ala 285 (301), Ala 306 (322), Ile 314 (330), Cys 315 (331), and Lys 501 (517). Asp 219 (235) and Ile

314 (330) line the top; Trp 223 (239), Thr 262 (278) and Ile 299 (283) the bottom; and Ala 220 (236), Ala 285 (301), Ala 306 (322) and Cys 315 (331) the sides of the PheA binding pocket. Residue Asp 219 is highly conserved through the A domains; it is invariant in amino acid activating A domains. From the crystal structure this residue was identified as being well positioned to form hydrogen bonds with the substrate α-amino group.

The tenth binding pocket residue, the strictly invariant Lys residue (501, pdb: 517) from the A10 motif is contributed by the A_{sub} domain and resides on a long loop that projects into the active site. This residue is well placed to form key polar interactions with both ligands; the α carboxyl group of the L-Phe substrate and the ribose O4′ and O5′ atoms of AMP.

One measure of ligand binding is an assessment of the hydrogen bonding between the ligand and protein. The average number of hydrogen bonds per nanosecond was calculated and will be used as a measure of the hydrogen bonding strength between particular residue groups.

Hydrogen bonding between ligands and PheA protein have been assessed from the average number of hydrogen bonds formed between each specified group of atoms per nanosecond versus time. In addition to this, the hydrogen bonding interaction of the Asp 219 and Lys 501 residues with the PheA have been analysed.

Analysis of these hydrogen bonding interactions and observation of the location of the ligands throughout the simulations reveals each ligand remains in its defined binding pocket on the timescale of the simulations, despite the relative motion of the A_{sub} domain. Several key differences are identified between the binding of the L-Phe substrate in the holo simulations. These differences will be discussed with reference to the difference in the principal modes of motion described by the essential eigenvectors.

Phenylalanine Substrate - Holo1

The hydrogen bonding interactions for the L-Phe ligand, Asp 219 (235) and Lys 501 (517) are shown in figure3.17.

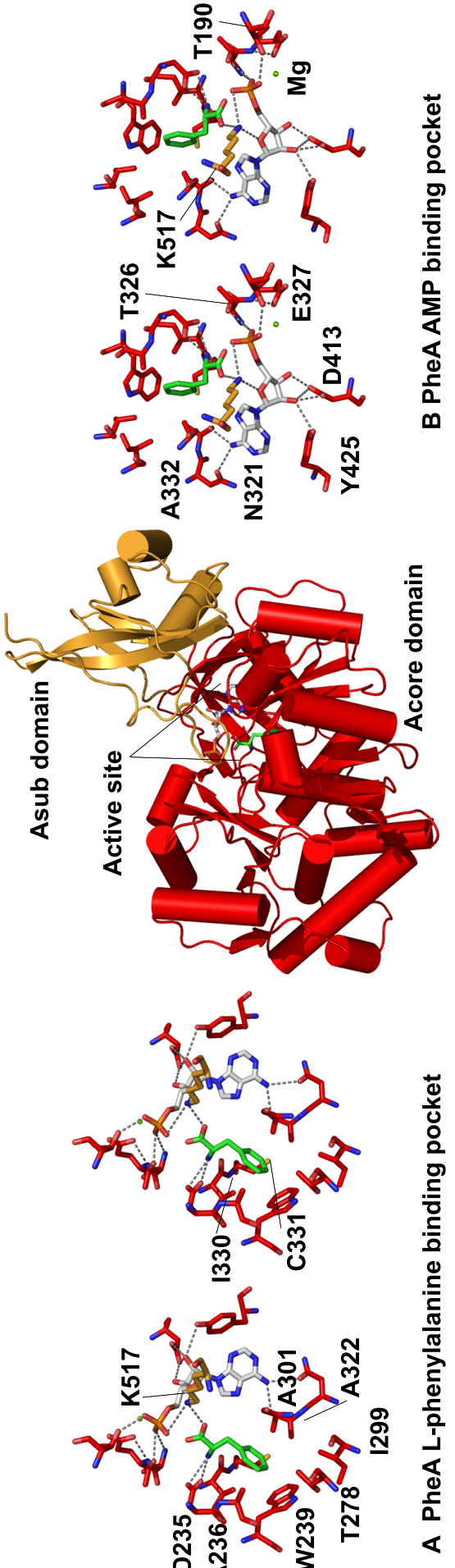


Figure 3.16: **The structure of the PheA.** A) The L-phenylalanine binding pocket showing the ten key ligand binding pockets in nred, AMP in grey, Phe in green and Mg in line. B) PheA Acore domain is in red and the A_{sub} domain in orange. The active site is formed at the interface between the two domains. C) The AMP binding site; key interaction residues in red[62].

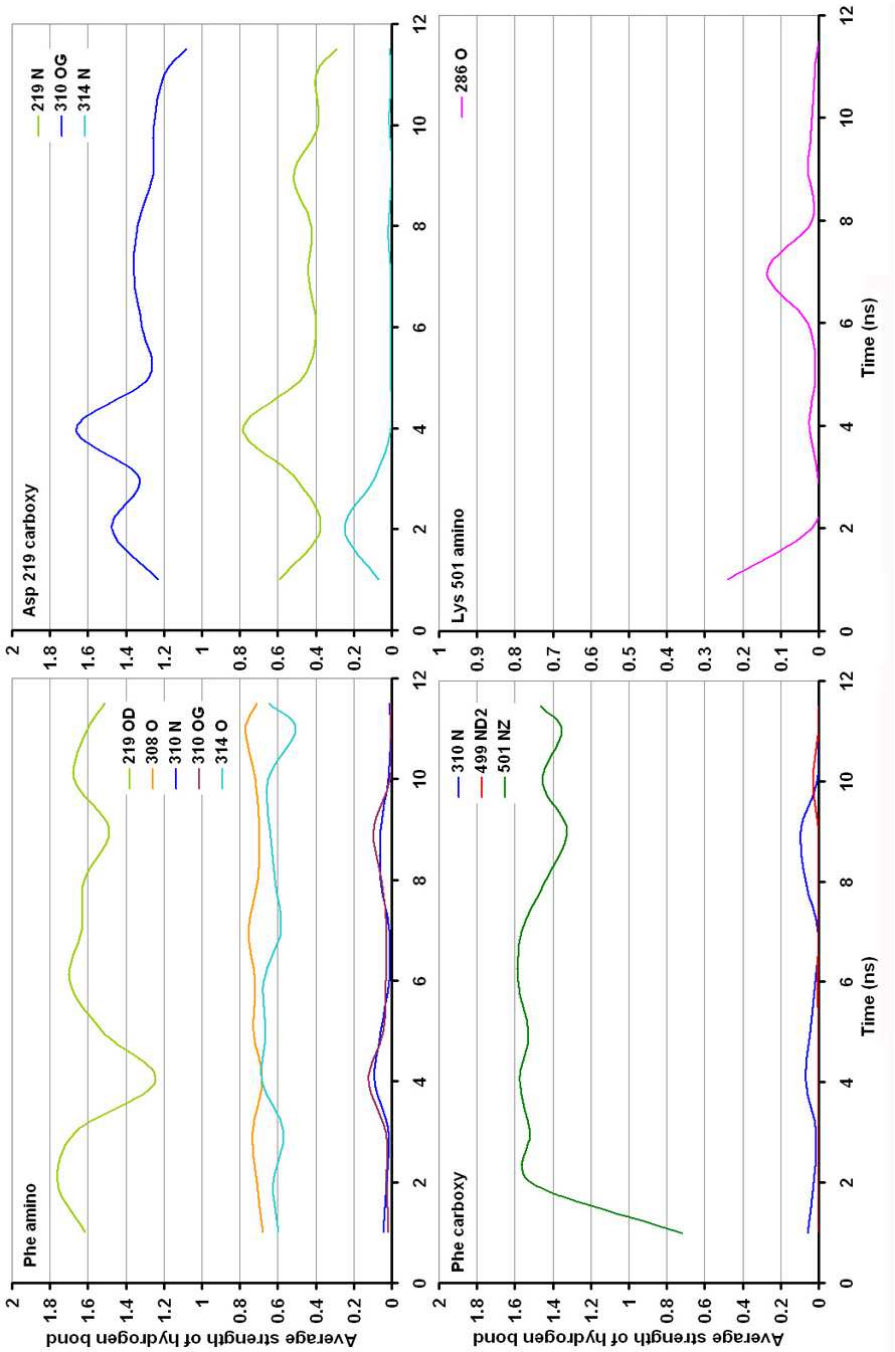


Figure 3.17: **Hydrogen bonding between the L-phenylalanine substrate and PheA in the PheA1-holo simulation.** The upper left graph shows the average strength of the hydrogen bond(s) formed between the Phe ligand amino group and the specified groups of PheA. The lower left graph shows the average strength of the hydrogen bond(s) formed between the Phe ligand carboxy group and the specified groups of PheA. The upper right graph shows the average strength of the hydrogen bond(s) formed between the Asp 219 PheA carboxy group and the specified groups of PheA. The lower right graph shows the average strength of the hydrogen bond(s) formed between the Lys 501 PheA amino group and the specified groups of PheA.

Strong hydrogen bonding, an average of 1.6, is observed between the α-amino group of the L-Phe substrate and the highly conserved Asp 219 residue (pdb: 235) throughout the PheA1-holo simulation.

Analysis of the hydrogen bonding of the Asp 219 carboxyl group with the PheA protein reveals that on the time scale of the simulation at least one hydrogen bond is maintained between this group and the side chain hydroxyl group of Thr 310. This bond is not observed in the PheA2-holo simulation. In the PheA2-holo simulation Thr 310 interacts with the phosphate moiety of AMP; an interaction not observed in the PheA1-holo simulation. An additional hydrogen bonding interaction of strength varying between 0.3–0.8 is formed between the Asp 219 carboxy group and the main chain amino group of Asp 219. The slight decrease in the strength of hydrogen bonding between the L-Phe substrate amino group and Asp 219 carboxyl group, observed during the fourth nanosecond, correlates with the slight increase in hydrogen bonding strength between the Asp 219 carboxyl group and the side chain hydroxyl group of Thr 310, and main chain amino group of Asp 219.

Hydrogen bonding is also observed between the α-amino group of the Phe substrate, and the main chain carbonyl group of Gly 308 and the main chain carbonyl group of Ile 314, bonding with an average strength of $\sim$0.7 and $\sim$0.6 respectively. The strength of the hydrogen bonding between the invariant Lys 501 residue and the carboxy group of the Phe substrate increases during the first 2 nanoseconds to $\sim$1.6. The strength of the hydrogen bonding interaction between these groups fluctuates between 1.3 and 1.6 during the remainder of the simulation.

Very few hydrogen bonding interactions are observed between the Lys 501 α-amino group and PheA. This is perhaps unsurprising given that the Lys 501 amino group forms key hydrogen bonds with both the L-Phe substrate and the AMP ligand.

Phenylalanine Substrate - Holo2

The pattern of hydrogen bonding interactions between the L-Phe substrate and PheA protein in the PheA2-holo simulation, as shown in the graphs on the left of figure 3.18, is notably

different to that observed in the PheA1-holo simulation.

In the PheA2-holo simulation, the strength of the hydrogen bonding interaction between the α-amino group of the L-Phe substrate and the carboxyl group of Asp 219 (pdb: 235) decreases from an average strength of $\sim$1.4 during the first 3 nanoseconds, to $\sim$0.15 during the sixth nanosecond, and rises gradually to 0.8 during the final 3.5 ns of the simulation. The reduction in the strength of this hydrogen bonding interaction is observed shortly after the time the extreme projection described by the first eigenvector is observed.

The 'stabilising' interaction between the Thr 310 side chain hydroxyl group and the Asp 219 carboxyl group, seen in the PheA1-holo simulation, is absent in the PheA2-holo simulation. A hydrogen bonding interaction is however observed between the side chain hydroxyl group of Thr 174 and carboxyl group of Asp 219. This interaction increases in strength as the interaction between the Asp 219 residue and L-Phe substrate amino group decreases. Interaction of the Asp 219 residue with a residue from the A3 motif loop is not observed in the PheA1-holo simulation where the flexibility of the A3 motif loop is greater.

The hydrogen bonding interaction between the L-Phe substrate α-amino group and the main chain carbonyl group of Gly 308 in the PheA2-holo simulation is similar to the interaction seen in the PheA1-holo simulation. The hydrogen bonding strength between the L-Phe substrate α-amino group and the main chain carbonyl group of Ile 314 fluctuates more on the time scale of the simulation, and is, on average, weaker in the PheA2-holo simulation.

The hydrogen bonding interaction between the Asp 219 carboxyl group and the main chain amino group of Asp 219 in the PheA2-holo simulation (see figure 3.18) is comparable to that observed in the PheA1-holo simulation. An additional strong hydrogen bonding interaction ($\sim$1) is formed between the main chain amino group of Ala 220 and the Asp 219 carboxyl group.

At least one hydrogen bond is present between the Lys 501 α-amino group and the L-Phe substrate carboxyl group during the simulation. The strength of this hydrogen bonding interaction fluctuates (between 1 and 1.6) more in the PheA2-holo simulation than in the PheA1-holo simulation. Additional hydrogen bonding is present between the L-Phe sub-

strate α-carboxyl group and PheA in PheA2-holo simulation that are not present in the PheA1-holo simulation. A fairly strong (0.8–1.35) hydrogen bonding interaction is formed between the main chain amino group of Gly 286 and L-Phe substrate α-carboxyl group between the fifth and eighth nanoseconds. An intermittent weaker (0.4–0.6) hydrogen bonding interaction is observed between the Phe substrate carboxyl group and Asn 499 side chain amino group during the PheA2-holo simulation.

The L-Phe binding pocket with the hydrogen bonding interactions at 0, 6 and 11.5 ns in each simulation is shown in figure 3.19). At the start the following interactions are observed:

- Thr 174 hydroxyl side chain with the phosphate group of the AMP ligand;
- Asp 219 α-carboxyl group with the α-amino group of the L-Phe substrate;
- Asp 219 α-carboxyl group with the main chain amino group of Asp 219, and
- Lys 501 amino group with the α-carboxyl group of the L-Phe substrate.

In the PheA2-holo simulation by 6 ns, Thr 174, located on the A3 motif loop, is in close proximity to the L-Phe binding pocket and the side chain hydroxyl group forms a hydrogen bonding interaction with the carboxyl group of Asp 219. This hydrogen bonding is accompanied by a slight alteration in the positioning of the L-Phe ligand in the binding pocket. The hydrogen bonding interaction between Asp 219 and Thr 174 persists until the end of the PheA2-holo simulation. The weaker interaction of Asp 219 with the L-Phe substrate appears to cause more flexibility in the positioning of the Phe substrate within the binding pocket. In comparison, the positioning of the Phe substrate, and original hydrogen bonding interactions between the substrate and PheA are preserved on the time scale of the PheA1-holo simulation.

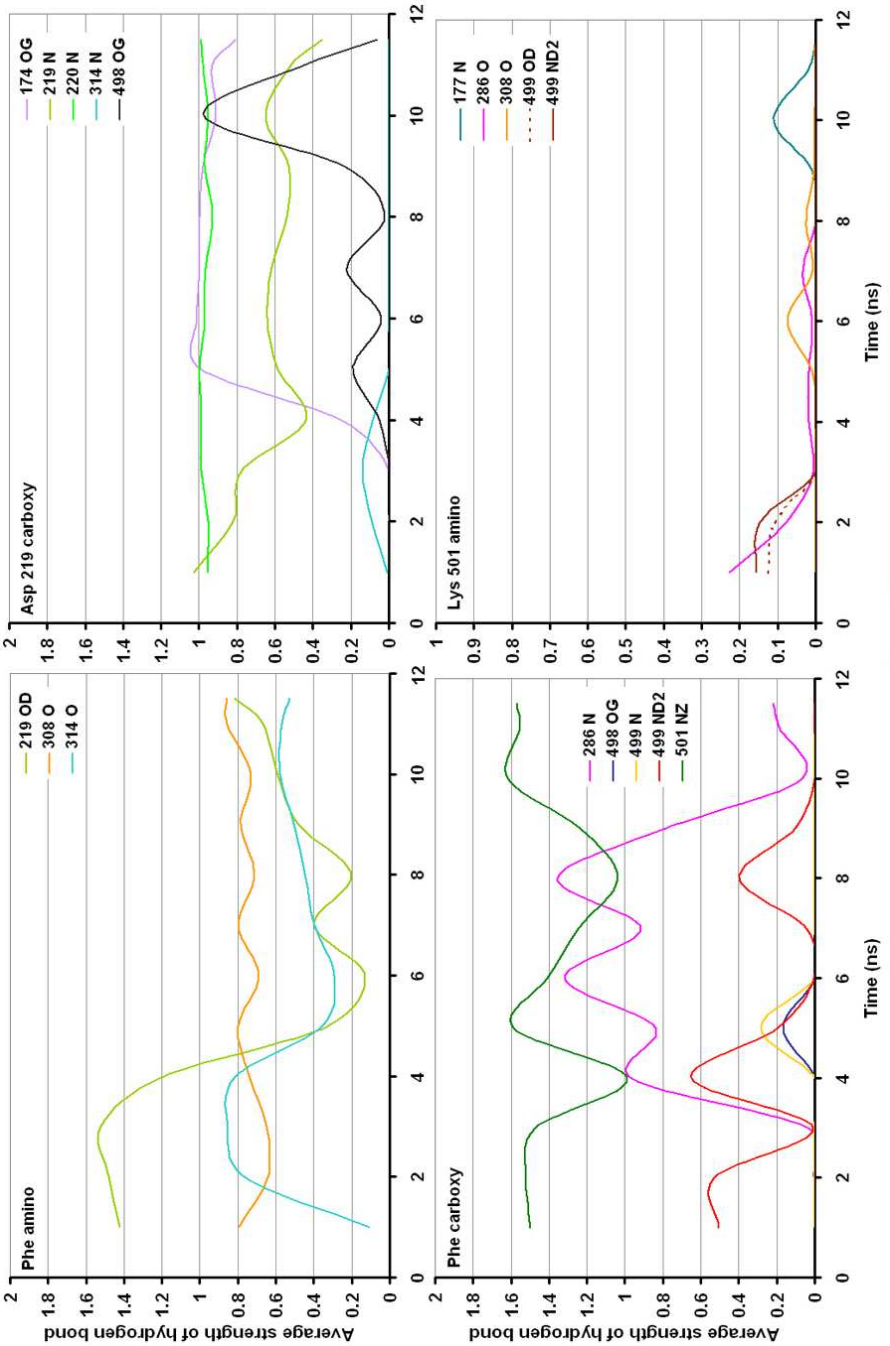


Figure 3.18: **Hydrogen bonding between the L-phenylalanine substrate and PheA in the PheA2-holo simulation.** The upper left graph shows the average strength of the hydrogen bond(s) formed between the Phe ligand amino group and the specified groups of PheA. The lower left graph shows the average strength of the hydrogen bond(s) formed between the Phe ligand carboxy group and the specified groups of PheA. The upper right graph shows the average strength of the hydrogen bond(s) formed between the Asp 219 PheA carboxy group and the specified groups of PheA. The lower right graph shows the average strength of the hydrogen bond(s) formed between the Lys 501 PheA amino group and the specified groups of PheA.

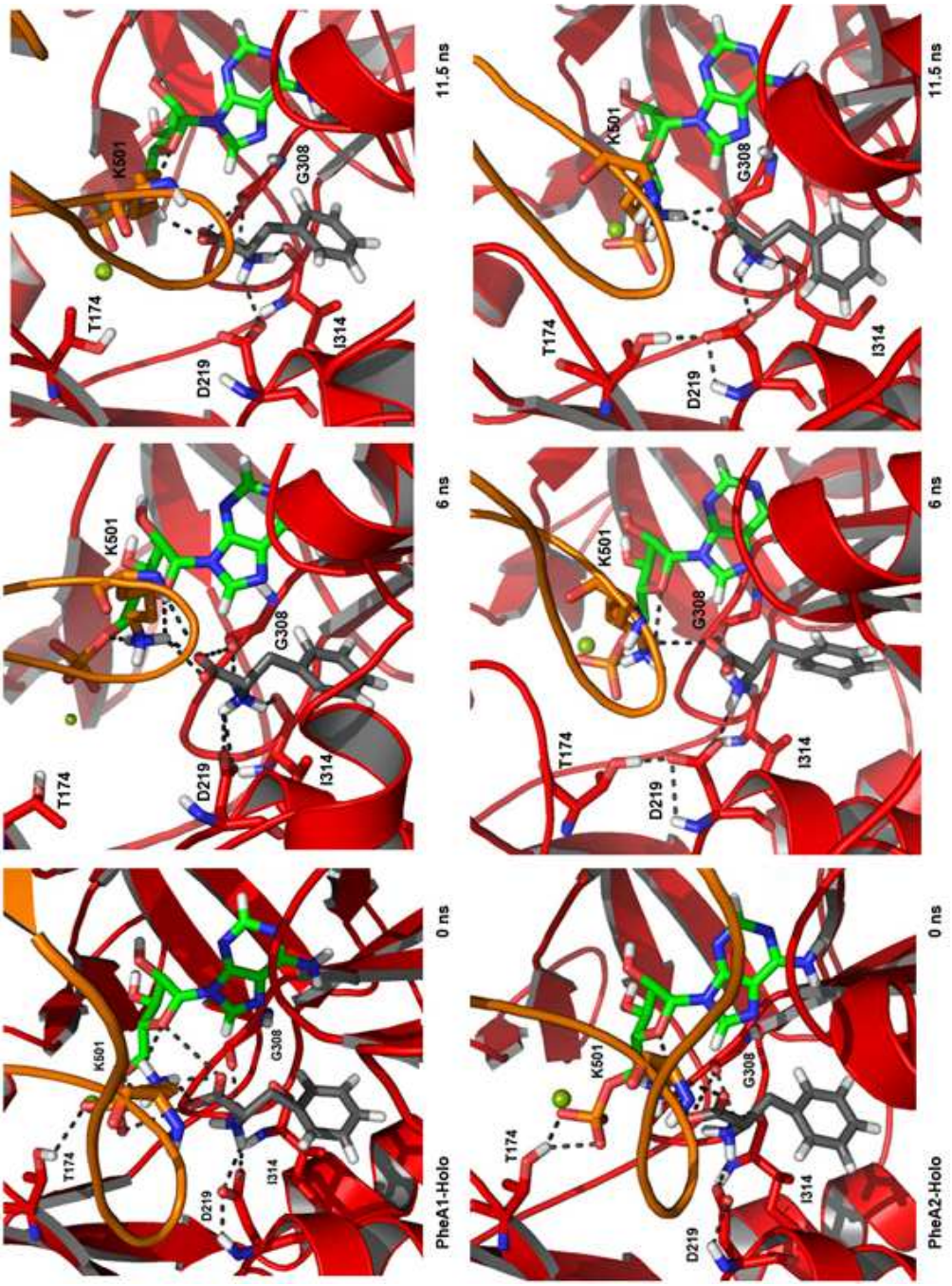


Figure 3.19: **Hydrogen bonding between the L-phenylalanine substrate and PheA in the PheA1- and PheA2-holo simulations at 0, 6 and 11.5 ns.** The upper set of three horizontal images shows snapshots from the PheA1-holo simulation of the PheA binding pocket at 0, 6 and 11.5 ns. The lower set of three horizontal images shows snapshots from the PheA2-holo simulation of the PheA binding pocket at 0, 6 and 11.5 ns.

3.3.14 AMP Binding

Adenine Binding

Key interactions identified by Conti *et. al*[62] between the adenine moiety of AMP and the PheA protein are between the exocyclic N6 nitrogen of the adenine group and the main chain carbonyl group of Ala 306 and the side chain oxygen atom of Asn 305. Hydrogen bonding of the protein to this nitrogen group is the major determinant by which the enzyme discriminates against guanine[62].

The pattern of hydrogen bonding between adenine and PheA in the PheA1-holo simulation versus time is shown in figure 3.20. In this simulation the strongest and most consistent hydrogen bonding interaction between the adenine and PheA is formed between the N6 group and the main chain carbonyl of Ala 306. A single hydrogen bond is formed between these atoms throughout the simulation. During the first three nanoseconds of the simulation the hydrogen bonding interaction between the side chain oxygen of Asn 305 and the N6 group of adenine is weak; 0.2–0.3, gradually increasing during the simulation to 0.9 in the final 1.5 ns of the simulation.

Hydrogen bonding interactions between PheA and the adenine moiety of AMP in the PheA2-holo simulation are shown in figure 3.21. The key hydrogen bonding interactions between the exocyclic N6 nitrogen of the adenine group and the main chain carbonyl group of Ala 306 and the side chain oxygen of Asn 305 are present in this simulation, however the interactions vary in strength throughout the simulation. The hydrogen bonding interaction between the N6 group and main chain carbonyl group of Ala 306 is of strength 1 for the first three nanoseconds of the simulation, weakening after this point to vary between a low of 0.6 during the eighth nanosecond of the simulation and 0.9 during the tenth nanosecond. As in the PheA1-holo simulation the hydrogen bonding interaction between the side chain oxygen of Asn 305 and the N6 group of adenine is weaker for the first three nanoseconds. For the remainder of the simulation the strength of this hydrogen bond varies between 0.7 and 0.9.

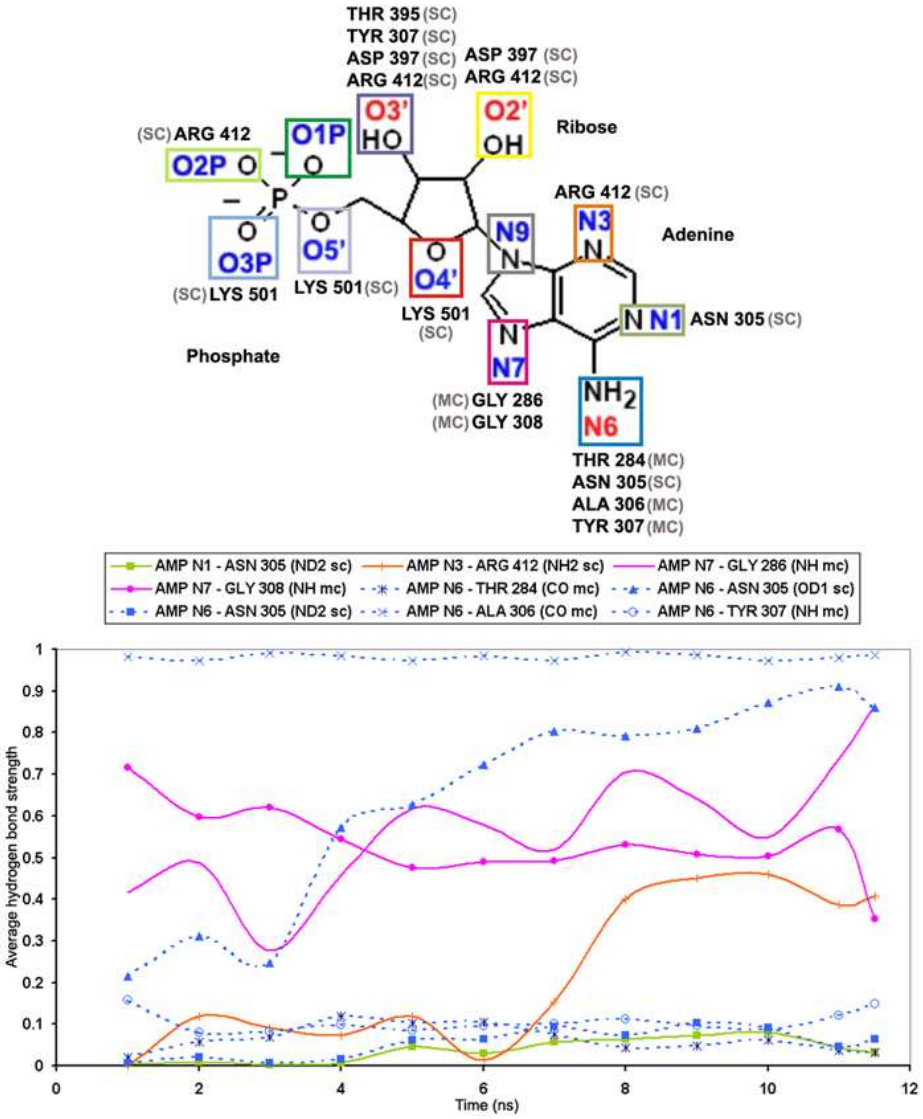


Figure 3.20: **Hydrogen bonding between the adenine moiety of the AMP ligand and PheA in the PheA1-holo simulation. Dashed lines represent interactions where the AMP atom / group is a donor, and solid lines represent interactions where AMP is an acceptor.**

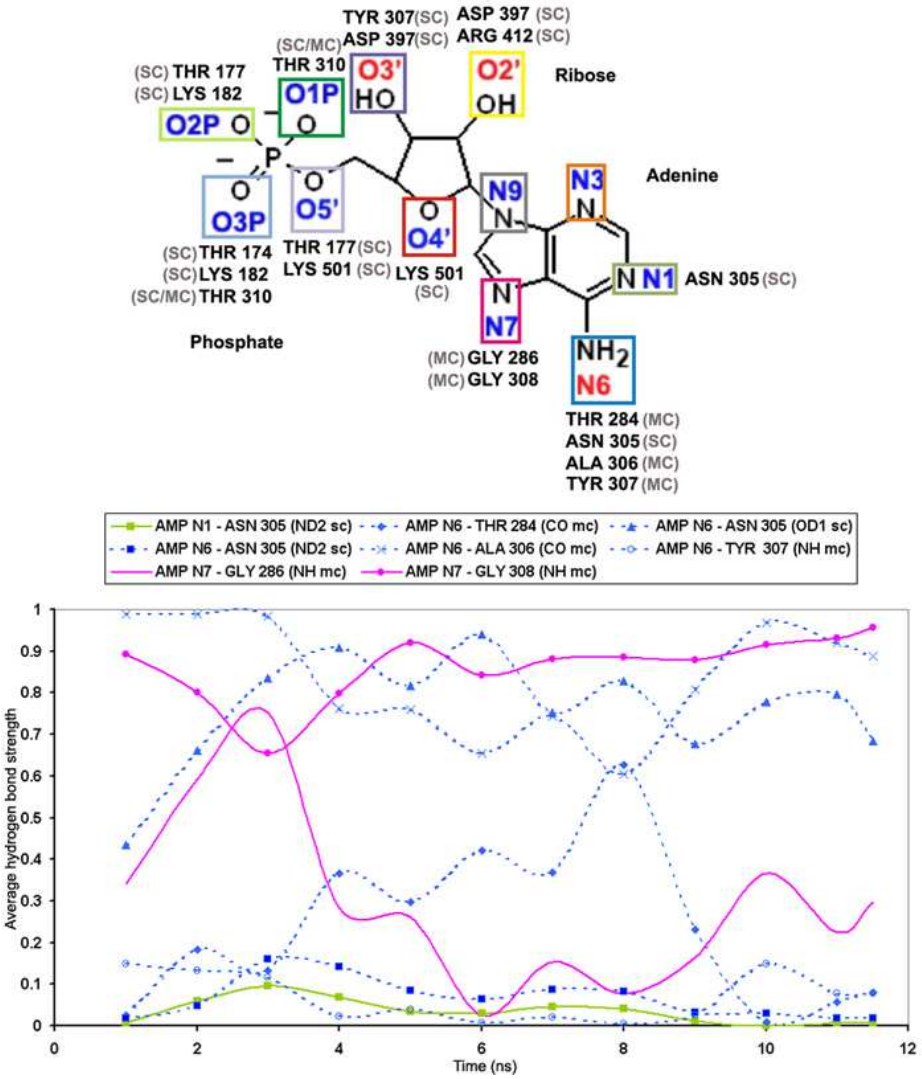


Figure 3.21: **Hydrogen bonding between the adenine moiety of the AMP ligand and PheA in the PheA2-holo simulation. Dashed lines represent interactions where the AMP atom / group is a donor, and solid lines represent interactions where AMP is an acceptor.**

In the PheA1-holo simulation a hydrogen bonding interaction with an average strength of 0.5 throughout the simulation, is formed between the N7 group of adenine and the main chain amino group of Gly 308. Overall this interaction is stronger in the PheA2-holo simulation; during the final 6.5 ns of this simulation this interaction is of strength 0.9. Conversely, the interaction between the N7 group of adenine and the main chain amino group of Gly 286 is stronger in the later stages of the PheA1-holo simulation than in the PheA2-holo simulation.

Ribose Binding

A number of interactions between the protein and ribose moiety of AMP were identified by Conti *et. al*[62] in the PheA crystal structure. These include:

- hydrogen bonds between invariant Asp 397 (pdb: 413) and the two hydroxyls of the sugar groups of the ribose moiety;
- a hydrogen bond between the ribose O4' and the invarient Lys 501 (pdb: 517);
- a long hydrogen bond between the 2' hydroxyl and the side chain of Tyr 409 (pdb: 425), and
- the close proximity of Tyr 307 (pdb: 323) to the adenine ring.

The hydrogen bonding interactions between PheA and the AMP ribose moiety in the PheA1-holo and PheA2-holo simulations are shown in the upper graph of figures 3.22 and 3.23, respectively.

In both holo simulations strong hydrogen bonding (1.6–1.9) is observed between the 2' hydroxyl of the ribose moiety and invariant Asp 397, and weaker hydrogen bonding (1.0–1.2) between the 3' hydroxyl of the ribose moiety and invariant Asp 397. In the PheA2-holo simulation the hydrogen bonding interaction between the 3' hydroxyl and Asp 397 is weaker for the first three nanoseconds of the simulation, starting at 0.2 and increasing to 1.0 by the end of the third nanosecond. Weak hydrogen bonding is observed in each simulation between the 3' hydroxyl atom and side chain hydroxyl of Tyr 307 and no hydrogen bonding is observed between the 2' hydroxyl and the side chain of Tyr 409 in either simulation.

Moderately strong (0.5–0.8) hydrogen bond is formed between the O4' atom of ribose and the amino group of the A10 motif lysine residue (Lys 501, pdb: 517) is observed in both holo simulations.

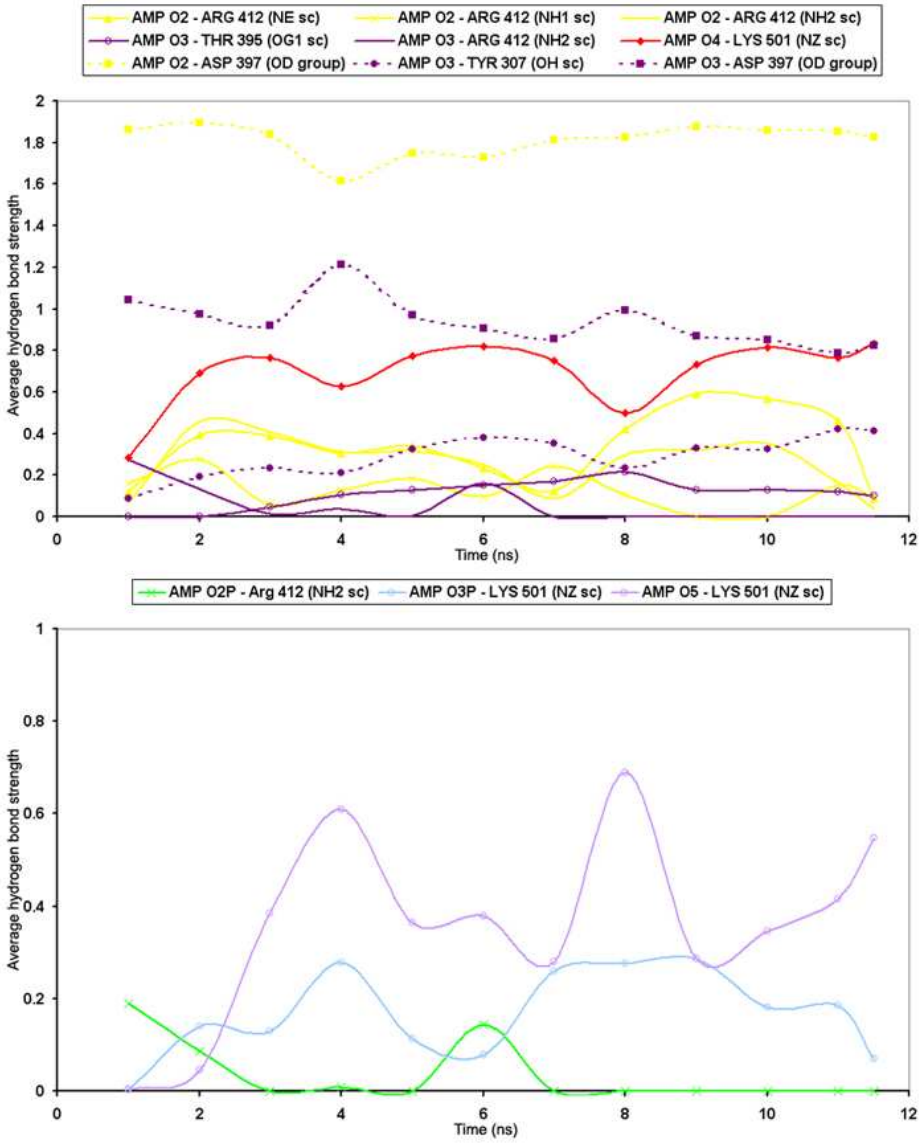


Figure 3.22: **Hydrogen bonding between the ribose (upper graph) and phosphate (lower graph) moieties of the AMP ligand and PheA in the PheA1-holo simulation. Dashed lines represent interactions where the AMP atom / group is a donor, and solid lines represent interactions where AMP is an acceptor.**

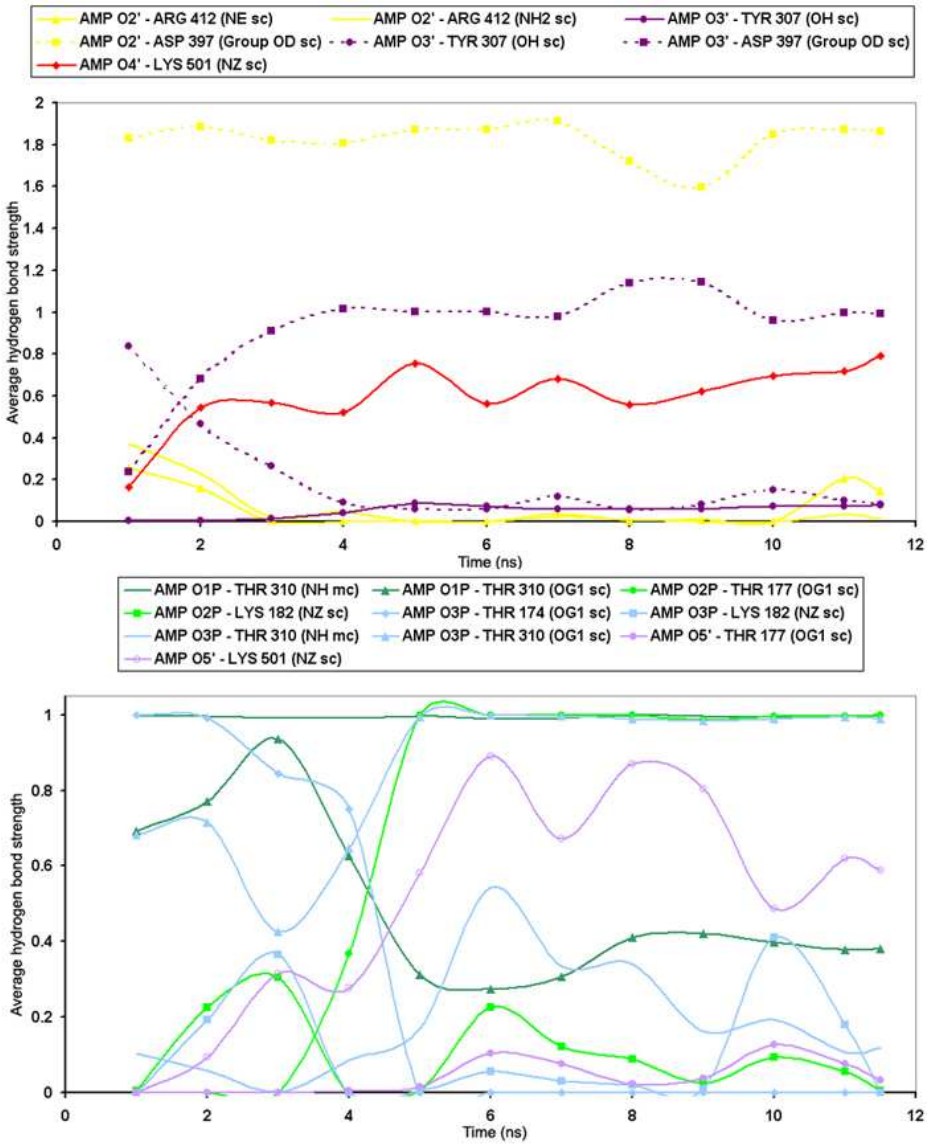


Figure 3.23: **Hydrogen bonding between the ribose (upper graph) and phosphate (lower graph) moieties of the AMP ligand and PheA in the PheA2-holo simulation. Dashed lines represent interactions where the AMP atom / group is a donor, and solid lines represent interactions where AMP is an acceptor.**

Phosphate Binding

The α-phosphate atom of AMP had slightly weaker electron density in the crystal structure indicating that the phosphate binding site is more disordered[62]. Key residues determined from the crystal structure to interact with this group include Glu 311 (pdb: 327), Thr 310 (pdb: 326) and Thr 174 (pdb: 190).

The hydrogen bonding patterns between the phosphate moiety of AMP and PheA are quite different in the PheA1- and PheA2-holo simulations. This is perhaps unsurprising given that this portion of AMP is in close proximity to the A3 motif loop which exhibits differing flexibility in the holo simulations.

Hydrogen bonding between the O5 atom and amino side chain of Lys 501 (pdb: 517) is common to both simulations. This hydrogen bonding fluctuates in strength throughout both simulations and is stronger in the PheA2-holo simulation.

Numerous hydrogen bonding interactions are present between PheA and the AMP phosphate moiety in the PheA2-holo simulation. Interactions are predominantly formed by residues in the A3 motif loop and Thr 310 (pdb: 326). Strong hydrogen bonding is observed between the O1 atom of the phosphate group and the amino main chain group of Thr 310; a single hydrogen bond is persistently present through the simulation. Weaker hydrogen bonding, $\sim$0.4 from 5 ns onwards, is observed between the O1 atom of the phosphate group and the hydroxyl side chain group (atom OG1) of Thr 310. This residue is highly conserved within the superfamily of adenylate forming enzymes.

The phosphate moiety of AMP forms very few hydrogen bonds with PheA in the PheA1-holo simulation. No hydrogen bonds are formed with Thr 310 or the residues of the A3 motif loop; which exhibits much greater flexibility in this simulation as compared with the PheA2-holo simulation. As outlined below, only one oxygen atom from the phosphate group is ligated to the Mg^{2+} in this simulation; in comparison two AMP phosphate oxygens which co-ordinate to the Mg^{2+} ion in the PheA2-holo simulation. As indicated by the experimental data, binding of the phosphate moiety is more disordered in both simulations than that of either the adenine or ribose moieties of AMP.

3.3.15 Mg Coordination

The position and ligands bound to the Mg^{2+} ion were analysed. In each of the PheA holo systems the positioning of th Mg^{2+} ion has moved from its starting position slightly to coordinate with six oxygen atoms in a distorted octahedral geometry. This is consistent with the preferred coordination number of divalent magnesium which is six.

In the PheA1-holo simulation the ligands to the Mg^{2+} are the carboxylate oxygen atoms of Glu 311 (pdb: 327), one oxygen of the AMP phosphate (O1P) and three water molecules (OW1, OW2 and OW3). In the PheA2-holo simulation the ligands are the carboxylate oxygen atoms of Glu 311, two oxygens of the AMP phosphate (O1P and O2P) and two water molecules (OW1 and OW2).

The mean and standard deviation for the bond lengths (Mg-O) and angles (O-Mg-O) were calculated, using data collected every picosecond, for the first and last ns, and over the entire simulation. These data for the PheA1- and PheA2-holo simulations are presented in Appendix 7.1.2 in figures 7.9, and 7.10 respectively.

Each mean bond length varies only slightly on the simulation timescales (the maximum standard deviation value is 0.008 nm for PheA1-holo and 0.007 nm for PheA2-holo). The bond angles (see in table 2 of the figures) exhibit a greater degree of variation in each system as reflected by the standard deviation value for each angle calculated over the entire simulation. Overall the standard deviation of each angle decreases throughout the simulation as indicated in the values from the first and last nanosecond.

3.4 Conclusions

In this chapter the results of MD simulations with the L-Phe GrsA A domain (PheA) from *Bacillus brevis*[62] totalling 46 ns are presented. These simulations were designed to explore the dynamics of the PheA A domain and understand the effect of the presence and absence of the hydrolysed products of the first half reaction on the dynamics of protein. To date no

molecular simulation study of the A domains has been reported in the literature.

The A domains belong to the Adenylate-forming superfamily. As discussed in section 1.4.3 of Chapter 1, "domain alternation" has been proposed as a strategy exploited by members of this superfamily to reconfigure the single active site of the enzyme to perform the two half reactions. Until 2012 A domain structures have only been determined in a first half reaction conformation and as the superfamily of enzymes share a conserved fold and highly conserved motifs, it has been proposed the A domains may also exploit a domain alternation strategy. Recently Mitchell *et al*[170] determined the structure of PA1221, a novel NRPS A domain and the associated PCP domain in the second half conformation providing evidence that the A domains of NRPSs also utilise two distinct conformations of the same protein to catalyse the two half reactions.

The first eigenvector of each holo simulation describes a rotation of the A_{sub} domain or a portion of the A_{sub} domain in a clockwise direction and tilting towards the left side of PheA.

In the PheA1-holo simulation the rotation occurs between the subdomain E and helix H6 of the A_{sub} domain about the axis defined by the residues from the A3, A8 and A10 motifs. The hinge region contains the residues from the A8 motif (417–426) located after the highly conserved Asp residue (414, pdb: 430). In PheA2-holo rotation of the entire A_{sub} domain occurs about an axis defined by the A8 motif residues. The hinge residues for this motion are 414–416 which include Arg 412 and Asp 414. An arginine residue equivalent to Arg 412 in PheA (pdb: 428) was identified by Dieckmann and co-workers from limited proteolysis of TycA as being a site of intrinsic flexibility, which decreased in the presence of the ligands[75,113].

Overlay of the structures equating to the extremes of motion in each simulation with a representative structure of the second half-reaction conformation (acetyl-CoA synthetase (bAcS) pdb 1PG4), to identify the PheA PPant binding site, shows that in both simulations the observed motion increases the distance between the A_{core} and A_{sub} domain on the side of the enzyme where the Ppant arm is expected to bind. The structures were also

overlayed with the structure of the modular NRPS SrfAC synthetase and the A domain of SrfAC fitted to PheA to indicate the relative positioning of the PCP domain. It should be noted that the PCP domain from this SrfAC synthetase is primed to interact with the downstream C domain from this NRPS module and therefore is positioned far from the A domain. However, the PCP domain is flexible and is thought to undergo conformation changes that would enable interaction with all of the required NRPS domains, which may be mediated by the flexible linker regions between individual NRPS domains. Together the proposed location of the PCP domain and the PPant ligand binding site indicates that the observed widening of the distance between the A_{core} and A_{sub} domains creates an opening through which the flexible PCP domain and phosphopanteinyl arm could access the active site of the enzyme.

Figure 3.24 shows this opening for the PheA1-holo simulation and figure 3.25 the opening for the PheA2-holo simulation.

The largest opening between the domains in the PheA1-holo simulation, from the extreme of motion from eigenvector 1, is observed at 7 ns. The extreme of motion described by the second eigenvector, observed at 9.8 ns, describes the A_{sub} domain tilting towards the right side of the A_{core} domain (away from the A3 motif loop) reducing the proposed PPant access opening.

In the PheA2-holo simulation the second eigenvector describes the rotation of subdomain E and part of helix H6 of the A_{sub} domain in a clockwise direction tilting slightly towards the right of the protein, however the overall motion brings this moving region of the A_{sub} domain closer to the A_{core} domain in a lid closing like motion. This motion, the extreme of which is observed at 8.378 ns, after that of eigenvector 1 which is observed at 2.779 ns, reduces the proposed PPant access opening on the right side of PheA slightly (figure 3.26), although not to the extent observed in the PheA1-holo simulation. Between the times that the extremes of this motion are observed, Thr 174 (pdb: 190) from the A3 motif loop forms a hydrogen bonding interaction with the Asp 219 (pdb; 235) key binding pocket residue. Interaction between Thr 174 and Asp 219 weakens the hydrogen bonding interaction observed between Asp 219 and the L-Phe substrate. The L-Phe substrate re-

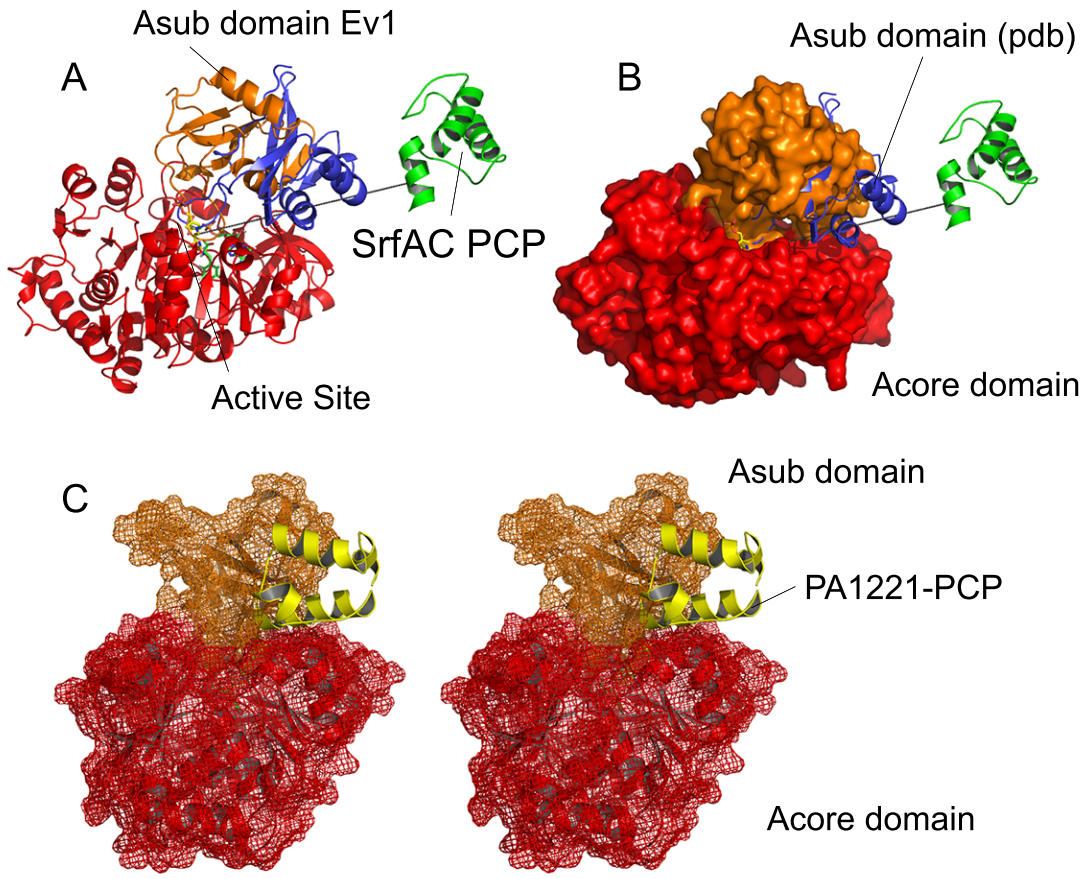


Figure 3.24: **Extreme motion described by eigenvector 1 in the PheA1-holo simulation.** A) The structure is shown with the A_{core} domain in red, and the A_{sub} domain in orange. The starting structure of PheA is shown in blue - the C-α atoms of the A_{core} domain of the starting structure was fitted to those of the structure corresponding to the extreme motion described by eignevector 1. The SrfAC A domain A_{core} domain was also fitted to the PheA A_{core} domain to give an indication of the positioning of the downstream PCP domain, primed to interact with the C domain, which is shown in green. The proposed site of the PPant ligand (yellow) was established by fitting the C-α atoms of the A_{core} domain of PheA with those from the bAcs adenylate forming domain (1PG4). The dotted line is used to indicate the direction through the opening observed between the A_{core} and A_{sub} domain of PheA through which the PPant arm of the PCP domain may access the PheA active site. B) The same structure is shown with a space filling representation to more clearly show the access to the ppant active site. c) The C-α atoms of the A_{core} domain of the PheA structure corresponding to the extreme motion of eigenvector were fitted to the C-α atoms of the A_{core} domain of PA1221, an recently determined A domain from *Pseudomonas aeruginosa*. This A domain was determined in the the second half reaction confirmation and in complex with the PCP domain which is directly interacting with the A domain.

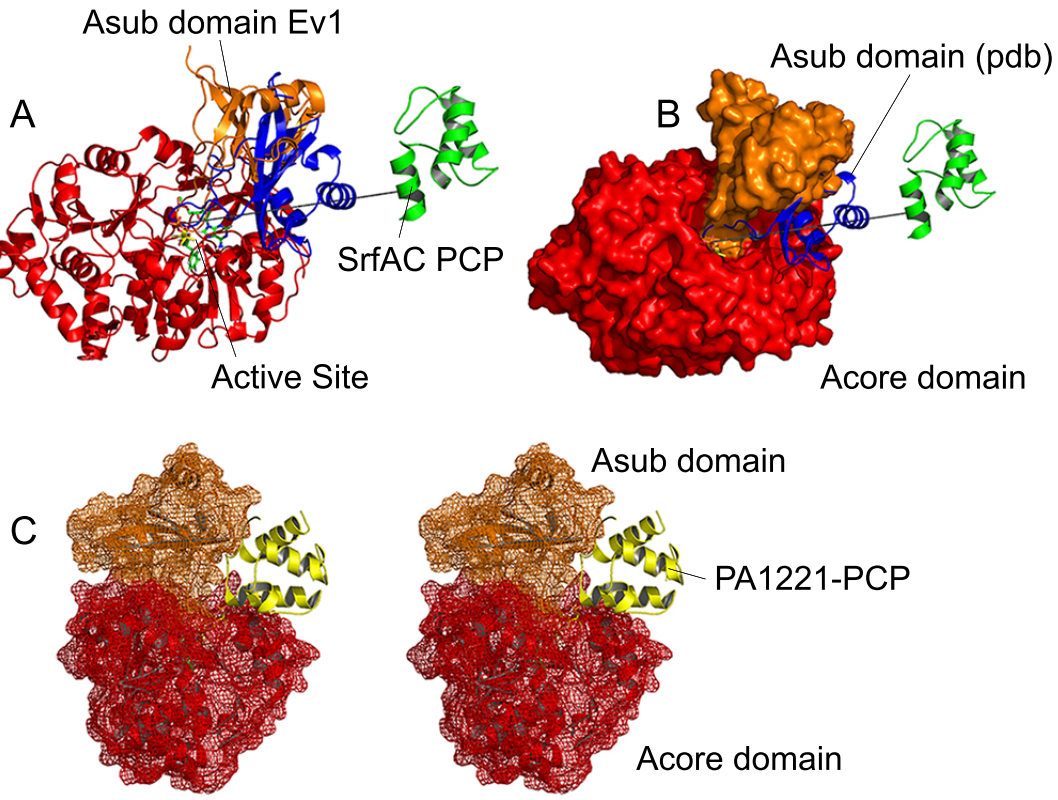


Figure 3.25: **Extreme motion described by eigenvector 1 in the PheA2-holo simulation.** A) The structure is shown with the A_{core} domain in red, and the A_{sub} domain in orange. The starting structure of PheA is shown in blue - the C-α atoms of the A_{core} domain of the starting structure was fitted to those of the structure corresponding to the extreme motion described by eignevector 1. The SrfAC A domain A_{core} domain was also fitted to the PheA A_{core} domain to give an indication of the positioning of the downstream PCP domain, primed to interact with the C domain, which is shown in green. The proposed site of the PPant ligand (yellow) was established by fitting the C-α atoms of the A_{core} domain of PheA with those from the bAcs adenylate forming domain (1PG4). The dotted line is used to indicate the direction through the opening observed between the A_{core} and A_{sub} domain of PheA through which the PPant arm of the PCP domain may access the PheA active site. B) The same structure is shown with a space filling representation to more clearly show the access to the ppant active site. c) The C-α atoms of the A_{core} domain of the PheA structure corresponding to the extreme motion of eigenvector were fitted to the C-α atoms of the A_{core} domain of PA1221, an recently determined A domain from *Pseudomonas aeruginosa*. This A domain was determined in the the second half reaction confirmation and in complex with the PCP domain which is directly interacting with the A domain.

mains in the binding pocket although it twists slightly. In this simulation the A3 motif is less flexible. Interaction between Asp 219 and Thr 174 is not observed in the PheA1-holo simulation, where the second largest mode of motion described a greater narrowing of the opening between domains on the right side of the protein and the A3 motif loop exhibits greater flexibility.

The interaction formed between Thr 174 from the A3 motif loop and Asp 219 may be required to maintain the opening between the A_{core} and A_{sub} domain through which the PPant arm may access the PheA active site, or this interaction may be an intermediate stabilising interaction required to facilitate further rotation of the A_{sub} domain. Interestingly, in the second half reaction conformation structure of CBL *A. sp. AL3007* His 207, which precedes the residue equivalent to PheA Asp 219, was shown to interact with a different residue from the A_{core} domain (Glu 310). This interaction, when compared with the structure of the CBL enzyme in the first half reaction, is shown to pull His 207 from the active site where it was occluding the PPant arm thiol from accessing the substrate[53].

The results from the PheA2-holo simulation suggest a role for the A3 motif loop in stabilising the enzyme to allow the second half reaction to take place.

3.4.1 Summary of Domain Motion

Figure 3.27 represents the motion observed between the A_{core} and A_{sub} domain in the PheA1-holo and PheA2-holo domains. In PheA-holo the largest motion occurs at 7ns with the A_{sub} domain moving towards the A3 motif loop and exposing and widening the Ppant active site. At 9.8ns the A_{sub} domain tips back towards the A3 motif loop reducing access to the Ppant active site. The A3 motif loop is flexible on the timescale of the simulation. In the PheA2-holo simulation the largest motion occurs at 2.8ns with the A_{sub} domain tipping towards the A3 motif loop exposing and widening the Ppant active site. As this happens residues from the A3 loop motif form interactions with residues from the Phe substrate binding pocket, pinning the A3 motif loop to the side of the PheA A_{core} domain and reducing the flexibility of this loop. These interactions reduce access to the left side of

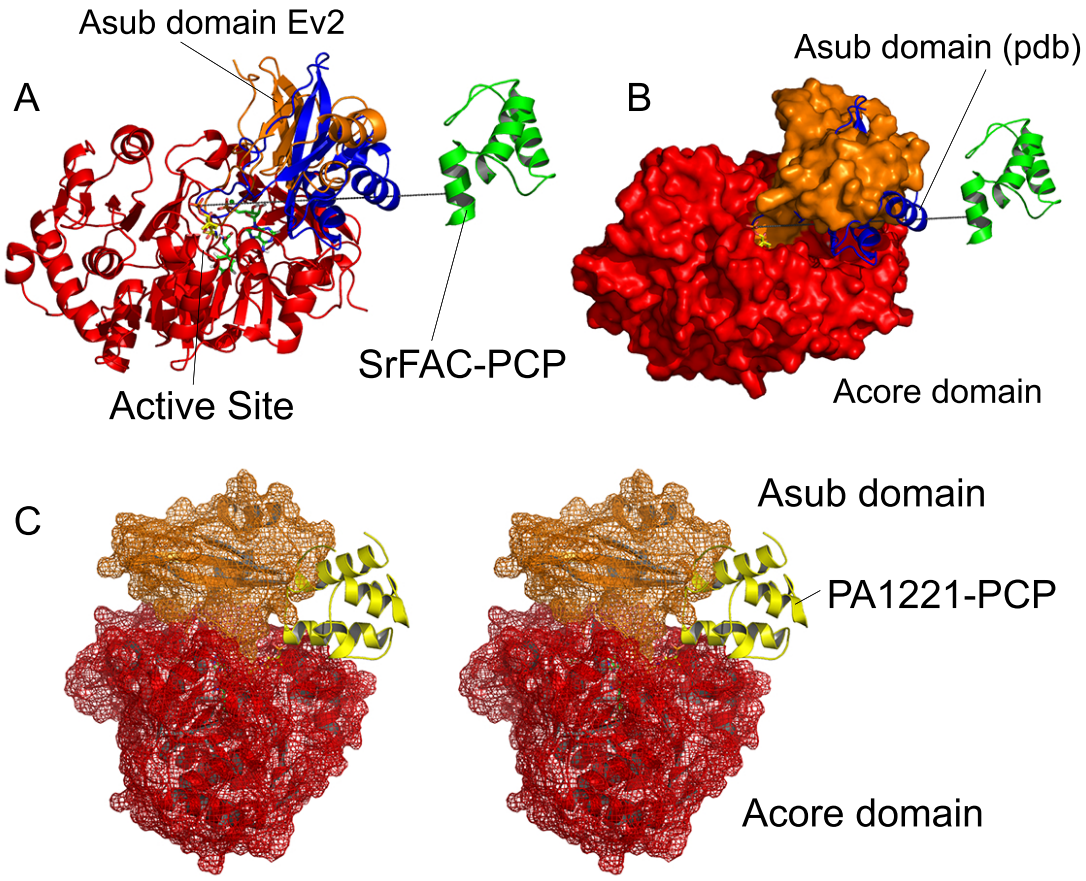


Figure 3.26: **Extreme motion described by eigenvector 2 in the PheA2-holo simulation.** A) The structure is shown with the A_{core} domain in red, and the A_{sub} domain in orange. The starting structure of PheA is shown in blue - the C-α atoms of the A_{core} domain of the starting structure was fitted to those of the structure corresponding to the extreme motion described by eignevector 1. The SrfAC A domain A_{core} domain was also fitted to the PheA A_{core} domain to give an indication of the positioning of the downstream PCP domain, primed to interact with the C domain, which is shown in green. The proposed site of the PPant ligand (yellow) was established by fitting the C-α atoms of the A_{core} domain of PheA with those from the bAcs adenylate forming domain (1PG4). The dotted line is used to indicate the direction through the opening observed between the A_{core} and A_{sub} domain of PheA through which the PPant arm of the PCP domain may access the PheA active site. B) The same structure is shown with a space filling representation to more clearly show the access to the ppant active site. c) The C-α atoms of the A_{core} domain of the PheA structure corresponding to the extreme motion of eigenvector were fitted to the C-α atoms of the A_{core} domain of PA1221, an recently determined A domain from *Pseudomonas aeruginosa*. This A domain was determined in the the second half reaction confirmation and in complex with the PCP domain which is directly interacting with the A domain.

the the A_{core} domain. The second greatest motion is observed at 8.4 ns where the A_{sub} domain has moved slightly away from the A3 loop motif and towards the Ppant active site but this does not greatly reduce access to the Ppant active site as the A_{sub} domain also moves downwards towards the A_{core} domain.

In Chapter 4 the results from a series of simulations of PheA with noncognate substrates is presented. These simulations were designed to observe effect of noncognate substrates on the dynamics of PheA including the behaviour of the A3 motif loop and to probe the role of the key hydrogen bonding interactions between the substrate and binding pocket resides and domain rotation.

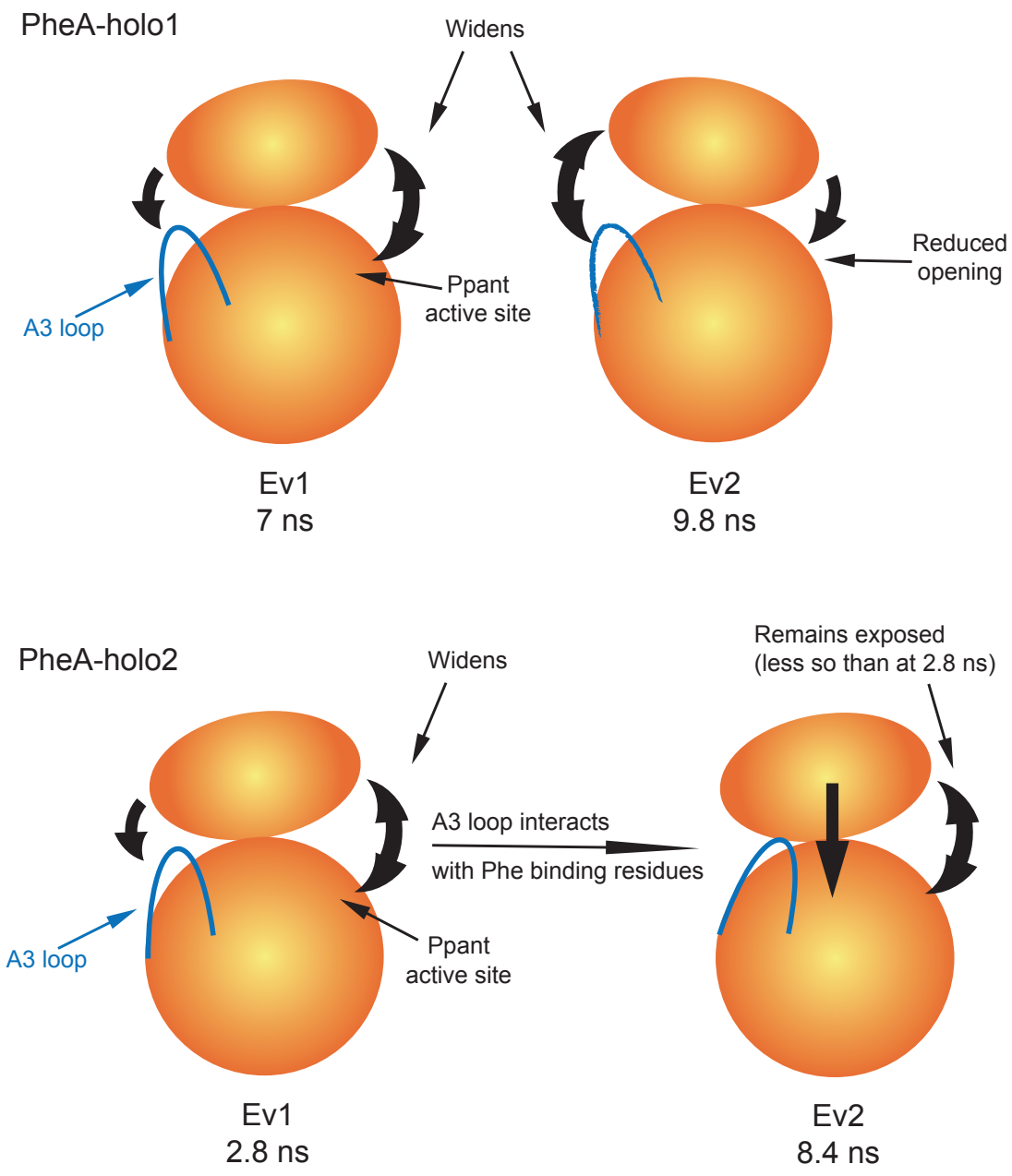


Figure 3.27: **Schematic of the domain motion observed in the PheA-holo simulations.**

Chapter 4

Molecular dynamics simulations of the phenylalanine activating adenylation domain, PheA, with Noncognate Substrates

4.1 Overview

In this chapter the results of a classical MD simulation study of the L-Phenylalanine activating gramicidin S synthetase (GrsA) A domain (PheA) from *Bacillus brevis*[62] with noncognate substrates is presented. Substrates of different size and physiochemical properties to L-Phe were selected for this study; L-Tyrosine, L-Arginine and L-Aspartic acid. In each simulation the noncognate ligand was docked into the L-Phe binding site. It must be acknowledged that in nature PheA would not naturally bind these substrates. These simulations were designed to observe the effect of the noncognate substrates on the dynamics of PheA and provide an understanding of which of the key residues are important for ligand recognition. These simulations indentify a potential role for the motif A3 loop in the removal of noncognate ligands from the PheA binding pocket and suggest that the substrate-Asp 219 (pdb: 235) and substrate-Lys 501 (pdb: 517) interactions are important for rotation/ motion of the A_{sub} domain relative to the A_{core} domain.

4.2 Introduction

Understanding the molecular basis of the substrate specificity and the dynamics of the A domains is crucial for enabling successful manipulation of these domains to engineer novel therapeutic agents. In the A domain the substrate, ATP and Mg^{2+} cofactors enter the active site and react to form a high energy acyl-adenylate (half-reaction 1). Following the release of pyrophosphate the substrate is covalently tethered to the terminal thiol of the Ppant arm of the PCP domain (half-reaction 2). As discussed in section 3.1 of Chapter 1, domain alternation, characterised by rotation of the A_{sub} domain to present different residues to the active site, has been proposed as a strategy to reconfigure the single active of the Adenylation superfamily enzymes for the catalysis of these two distinct half-reactions[52,53,91]. The interactions which take place across the NRPS domains require a degree of flexibility in the PCP domain and, very likely, in other NRPS domains.

One strategy for producing novel antibiotics and molecules with pharmacologically attrac-

tive properties is the modification of the assembly line machinery of the nonribosomal peptide synthetase (NRPS). This can be accomplished by either altering or switching individual yet equivalent domains, or by exchanging entire modules. As the Adenylation (A) domains of NRPS are the primary, yet not exclusive, determinants of substrate selectivity they can be targeted to alter the substrates incorporated into the peptide product. Point mutation of the key ligand binding residues of the A domain, designed to switch the substrate selectivity preference, represents a less intrusive approach to NRPS engineering than excising and swapping entire A domains or modules. Both of the A domain altering strategies have implications for the recognition of the substrate by the downstream C domain. In addition the excision and replacement of individual NRPS domains may have implications on the 1–15 stretch of residues linking the domains, whose structural and functional role, if any, is not yet fully understood.

Bioinformatics sequence analysis studies of the A domains have been performed[81,82]. These studies have conferred a relationship between the residues lining the substrate binding pocket and the substrate the enzyme preferentially binds. Few structural modelling studies of the A domains have been performed and these have been limited to mutating the PheA binding pocket residues to visualise the structure of alternative substrate specific A domains[159] and building homology models into which the natural amino acid substrate has been docked[279]. No molecular simulation study of the Adenylation domains has been reported in the literature.

Some kinetic data are available for PheA and the noncognate ligands L-Tyr, L-Asp and L-Arg. In 2001 Luo *et al*, determined the dissociation constants (Kd) of various amino acids to the PheA adenylation domain of PheA using the equilibrium fluorescence titration method. With the exception of L-Arg all L-amino acids tested had a dissociation constant within 2-3-fold of L-Phe Kd suggesting that the phenylalanine-binding pocket is large enough to accommodate most other amino acids and not very sensitive to the size or charge of the substrate side chain[1].

The amino acid dependent ATP-PPi exchange assay was then used to assess apoPheATE-catalyzed aminoacyl adenylate formation and the continuous spectrophotometric pyrophos-

Amino Acid	K_d /μM
L-Phe	6 + - 1
D-Phe	7 + - 1
L-Tyr	2.0 + - 0.3
L-Arg	56 + - 15
L-Asp	3.3 + - 0.7

Table 4.1: Disocciation constants for binding of various amino acids to the adenylation domain PheA. Table adapted from Luo *et al.*[1].

phate assay used to measure amino acid-dependent ATP consumption.

4.3 Methods

4.3.1 System Preparation

The PheA structure with the modelled A3 motif loop, see section 3.2.1 in Chapter 3, was used as a starting structure for each simulation presented in this chapter. The addition of hydrogen atoms to the PheA is outlined in section 3.2.2.

Energy minimisation, performed using GROMACS 3.2.1, was used to relieve steric conflicts generated during the simulation setup. The AMP force field parameters, described in section 3.2.7, were used in all calculations performed using GROMACS. The convergence criteria for energy minimisation, *g = 0 +/- e*, is when the gradient (*g*) reaches a value within *e* of *0*. Unless otherwise specified minimisation was performed until either; *e* reached 1000 kJ mol^{-1} nm^{-1}, or the specified number of steps had been completed. During energy minimisation no constraints were placed on the bond lengths. Unless otherwise specified, when heavy atoms were tethered a harmonic potential with a force constant of 1000 kJ mol^{-1} nm^{-2} was used. Unrestrained energy minimisation is where no atoms were tethered.

4.3.2 Docking

Initial attempts were made to dock the ligands into the binding pocket of PheA in the 'apo' state. These docking runs proved unsuccessful, with the ligand binding in the enzyme active

	PheATE			apoPheATE		
Substrate	kcat (min-1)	Km (mM)	kcat/Km (min-1 mM-1)	kcat (min-1)	Km (mM)	kcat/Km (min-1 mM-1)
L-Phe	0.06 + - 0.01(a)	0.03 + - 0.01(a)	2a	690	0.07	9900
	0.07 + - 0.02 (b)	0.03 + - 0.02(b)	2.3b			
D-Phe	0.06 + - 0.01(a)	0.02 + - 0.008(a)	3a	720	0.07	10300
L-Trp	0.16 + - 0.02(a)	0.3 + - 0.1(a)	0.5a	552	0.74	750
L-Tyr	0.17 + - 0.01(a)	1.1 + - 0.1(a)	0.15a	10	0.31	32
	0.2 + - 0.02(b)	1.4 + - 0.1(b)	0.14b			

Table 4.2: **Kinetic Constants for Amino Acid-Dependent ATP Hydrolysis by ApoPheATE and HoloPheATE Measured by Continuous Spectrophotometric Pyrophosphate Assay and ATP-PPi Exchange Assay** a) Kinetic constants measured for amino acid-dependent ATP hydrolysis by apoPheATE. b) Kinetic constants measured for amino acid-dependent ATP hydrolysis by holoPheATE. Table from Luo *et al*[1].

site but in a conformation that prevented the AMP molecule docking into the binding site defined in the PheA crystal structure. The substrates were therefore, docked in to the PheA structure with the co-factors present.

The system - PheA, AMP and Mg^{2+} - was minimised using up to 100 steps of steepest descents during which all heavy atoms were tethered. This was followed by up to 500 steps of steepest descents energy minimisation where e = 10 and all heavy atoms except the magnesium ion were restrained. To prepare the files for docking, atoms were renamed as required to conform to the AutoDock naming convention. The AutoDock 3.0.5 program[220] with the Lamarckian genetic algorithm (LGA) was used for the docking simulations.

The AutoDockTools (ADT) program was used to prepare the structure (merge hydrogen atoms, add charges and solvation parameters), the docking grid and docking parameter files for input to AutoDock.

Non-polar hydrogen atoms were merged and Kollman charges and solvation parameters were added to PheA. Gasteiger charges were calculated for the AMP molecule and then the non-polar hydrogen atoms were merged. Solvation charges for AMP were obtained from the sol_par.py of ADT. A charge of +2 and a solvation parameter of 0 were used for the magnesium ion. Additionally, the phosphorous and magnesium ions were renamed to X and M, respectively.

Non-polar hydrogen atoms and Kollman charges were added to the ligand (either L-Tyr, L-Asp or L-Arg) and the rigid root of the ligand, and the number of active torsions determined. The docking grid was defined as to contain 90 x 90 x 90 points with a grid spacing of 0.275$\mathring{A}$. The grid was placed symmetrically about the centre of mass of the macromolecule. This ensured the grid boxes included the entire enzyme binding site and also provided enough space for the ligand translational and rotational walk. The appropriate parameters for atoms P (X) and Mg (M) as defined in the AutoDock documentation were added manually after the parameter file was written by the AutoGrid routine.

Docking Genetic Algorithm Parameters

For each PheA-ligand complex 100 runs were performed. For each run, a maximum number of 25,000 genetic algorithm (GA) operations was performed on a population of 50 individuals. The maximum number of energy evaluations was set to 250,000. The ligand was placed in a random starting position and conformation at the beginning of each docking run. During the docking simulation the ligand was allowed a maximum mutation of 0.2 Å in translation and 50 ° in rotation. The mutation rate was set to 0.02, the crossover rate to 0.8, elitism to 1 and the local search rate to 0.06 individuals in the population.

Evaluation of Docking Results

The docking simulation results were ranked according to the energy between the protein and the docked ligand - a summation of internal ligand energy and intermolecular energy terms. A conformational clustering analysis was performed on the resulting structures. The orientation and position of the top ranking structure of the ligand from each cluster was visualised in PheA using Visual Molecular Dynamics (VMD)[280] and compared to that of L-Phe in the crystal structure. The docked ligand which had an orientation most consistent with L-Phe in the PheA crystal structure was selected to be used as the starting structure for the MD simulations.

4.3.3 Energy Minimisation Protocol prior to Simulations

The PheA-Tyr, -Asp and -Arg systems were subjected to up to 100 steps of steepest descent minimisation with all heavy atoms tethered to their original position. After the addition of solvent (water and counterions), up to 100 steps of steepest descents minimisation was performed with all heavy atoms tethered to their original position. Following this, 100 steps of conjugant gradients minimization were performed with only the heavy atoms of the substrate (either Tyr, Asp or Arg) tethered using a harmonic potential with a force constant of 500 kJ mol^{-1} nm^{-2} . This was followed by up to a further 50 steps of steepest descents

and up to 50 steps of conjugant gradients unrestrained minimization.

4.3.4 Simulation Preparation

All simulations were performed in a truncated octahedral box, 770 nm^3, and the GROMACS genbox routine was used to solvate the systems. This routine fills the box with multiple translational images of a single configuration of 216 simple point charge (SPC)[269] water molecules, then removes water molecules when the distance between any atom of the solute molecule (protein or protein ligand complex) and any atom of the solvent molecule is less than the sum of the van der Waals radii of both atoms. An appropriate number of randomly selected water molecules was replaced with Na^+ ions using the genion GROMACS utility to achieve overall neutrality of each system. The resulting system sizes are listed in table 6.1.

Simulation	PheA-Tyr	PheA-Asp	PheA-Arg
Protein atoms	5213	5213	5213
Counterions (Na^+)	16	17	15
Water molecules	22935	22937	22936
Total atoms	74085	74083	74086

Table 4.3: Summary of the noncognate simulation systems.

After minimisation, each system was simulated in the canonical ensemble (NVT) with heavy atoms tethered to help to ensure relaxation of the solvent. NVT MD simulation for 250 ps, in which an isotropic force constant of 1000 kJ/mol^{-1} nm^{-1} was applied to tether all non-hydrogen atoms, was followed by a further 250 ps NVT MD simulation in which an isotropic force constant of 500 kJ/mol^{-1} nm^{-1} was applied to tether all heavy atoms.

Subsequent to this an un-tethered production run of 11.5 ns in the isothermal-isobaric ensemble was performed using the protocol outlined in section 3.2.6 of Chapter 3.

All simulations were performed on University of Warwick Centre for Scientific Computing Argus task farm between April 2005 and December 2006. The source code for GROMACS 3.2.1 was compiled by the author of this thesis. Each simulation was run on a single node; the approximate run time for a 1 ns simulation was 180 hours or 7.5 days. Each noncognate

ligand simulation took approximately 2160 hours to complete.

4.3.5 MD Simulation Analysis Methods

Analysis of the trajectories was carried out using the methods outlined in section 3.2.8, Chapter 3. In summary the following aspects of the system were analysed to provide an understanding of the protein dynamics and ligand binding:

- RMSDs and RMSFs of the protein Cα atoms;
- Secondary structure analysis as a function of time using DSSP criteria[273];
- Principal motion of PheA using principal components analysis and DynDom[274,275];
- Intramolecular and interdomain hydrogen bonding;
- Radius of gyration;
- Substrate and AMP cofactor hydrogen bonding interactions with PheA, and
- Mg-ligand coordination geometry.

4.4 Results and Discussion

The analysis of the classical MD simulations of PheA with the L-Tyr, L-Asp and L-Arg substrates is presented in this section. As for the simulations presented in Chapter 3, results for the entire production run of 11.5 ns have been analysed and are presented. These observations are discussed in detail in relation to the biological relevance and methods used. Where appropriate, these results will be compared with those from the PheA-apo and holo simulations.

4.4.1 Docking Results

Observations from the A domain sequence and structural data and the results of simulations carried out with PheA (as discussed in Chapter 3) indicate that the Asp 235 (pdb:219) and Lys 501 (pdb:517) residues at the top of the A domain binding pocket form electrostatic stabilising interactions with the substrate α-amino and α-carboxylate atoms. These residues are highly conserved and invariant, respectively, within the A domains. The existence of these interactions between the substrate and binding pocket, knowledge of the location of the substrate binding pocket and the residues within it, were used to help guide and assess the docking simulations.

Figure 4.1 shows the conformations of the L-Tyr, L-Asp, and L-Arg substrates selected from the docking simulations, within the active site of PheA, and used to initiate the relevant MD simulation.

The highest ranking structure of the L-Tyr ligand was obtained from run 34 of the docking simulation. In this conformation the L-Tyr substrate α-amino and α-carboxyl groups are appropriately orientated towards the Asp 219 and Lys 501 residues of PheA respectively, and the sidechain is well positioned within the L-Phe substrate active site.

The highest ranking structure (run 7) from the fourth highest ranking cluster of docked Asp structures was used as the starting structure for the PheA-Asp simulation. As for L-Tyr, this was the highest ranking structure that made the required substrate α-amino-Asp 219, and substrate α-carboxyl-Lys 501 interactions.

None of the docked structures of L-Arg made the appropriate substrate amino - Asp 219, and substrate carboxyl - Lys 501 interactions, this was not unexpected given the size of the substrate in comparison to the L-Phe substrate and the volume of the binding pocket. The highest ranked, lowest energy structure was used to initiate the MD.

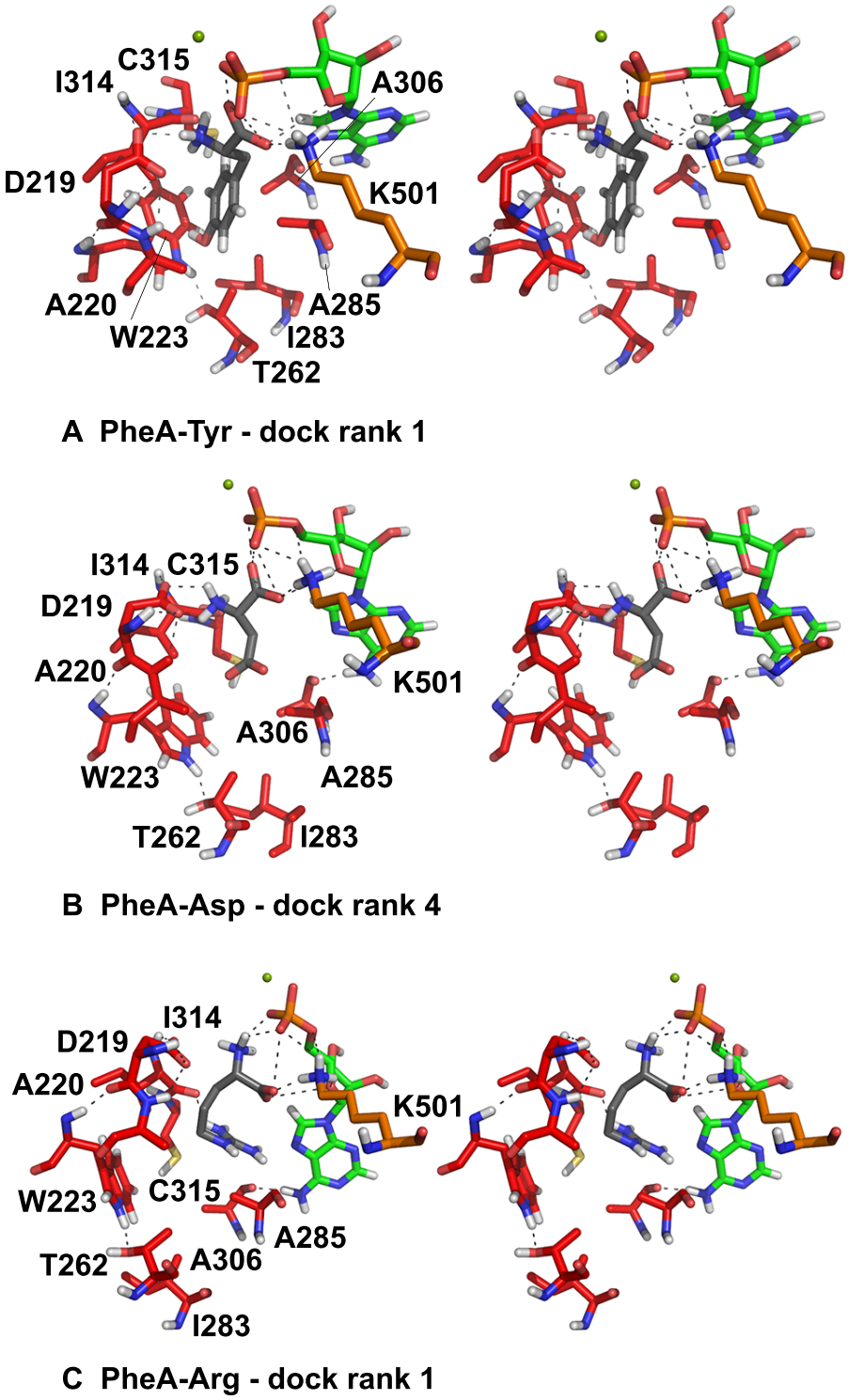


Figure 4.1: **Docked structures - PheA with L-Tyrosine, Aspartic acid and Arginine.** The confirmations of the PheA, and A) L-tyrosine, B) aspartic acid, and C) arginine complexes obtained from docking simulations and used to initiate the MD simulations PheA-Tyr, PheA-Asp, PheA-Arg, respectively. Substrate is shown in grey, AMP in green, Mg in lime and 10 key substrate binding pocket residues in red.

4.4.2 Global Structural Stability

The RMSD of PheA from its corresponding starting structures, after least-squares fitting (rigid body rotation and translation), was calculated to obtain information regarding the conformational stability of the protein on the timescale of the simulation. The RMSD for all Cα atoms, the Cα atoms in the A_{core} domain and A_{sub} domain was calculated as is shown in the upper graph of each relevant figure. As for the simulations presented in Chapter 3, the individual domains were decomposed into the linker region, secondary structural elements (helices and sheets), and loops, and the RMSD of these regions from the starting structure calculated. The RMSD of these regions in the A_{core} and A_{sub} domain is shown in the middle and lower graphs respectively of the relevant figures.

PheA-Tyr

Figure 4.2 shows the RMSDs calculated for the Cα atoms of PheA in the PheA-Tyr simulation.

The arching trend of the RMSD of all the Cα atoms observed in the PheA-holo simulations is not observed in the PheA-Tyr simulation. Comparison of the RMSD for the whole protein with that of the individual domains suggests there may be some relative domain motion in the simulation of PheA with L-Tyr.

The RMSD of the all Cα atoms fluctuates throughout the simulation rising and falling repeatedly. As observed in the PheA-holo simulations, the A_{core} domain RMSD reveals the domain to be stable on the timescale of the simulation. The A_{sub} domain, while more stable than in the holo simulations, still exhibits greater structural drift than the A_{core} domain.

The structural drift of the individual components of the A_{core} and A_{sub} domains are very similar to those of the holo simulations. The A_{core} sheets and helices are the most stable element of this domain with the loops and then the N-terminal linker region exhibiting greater structural drift.

The A_{sub} domain helices and sheets RMSD indicate this domain is slightly more stable in

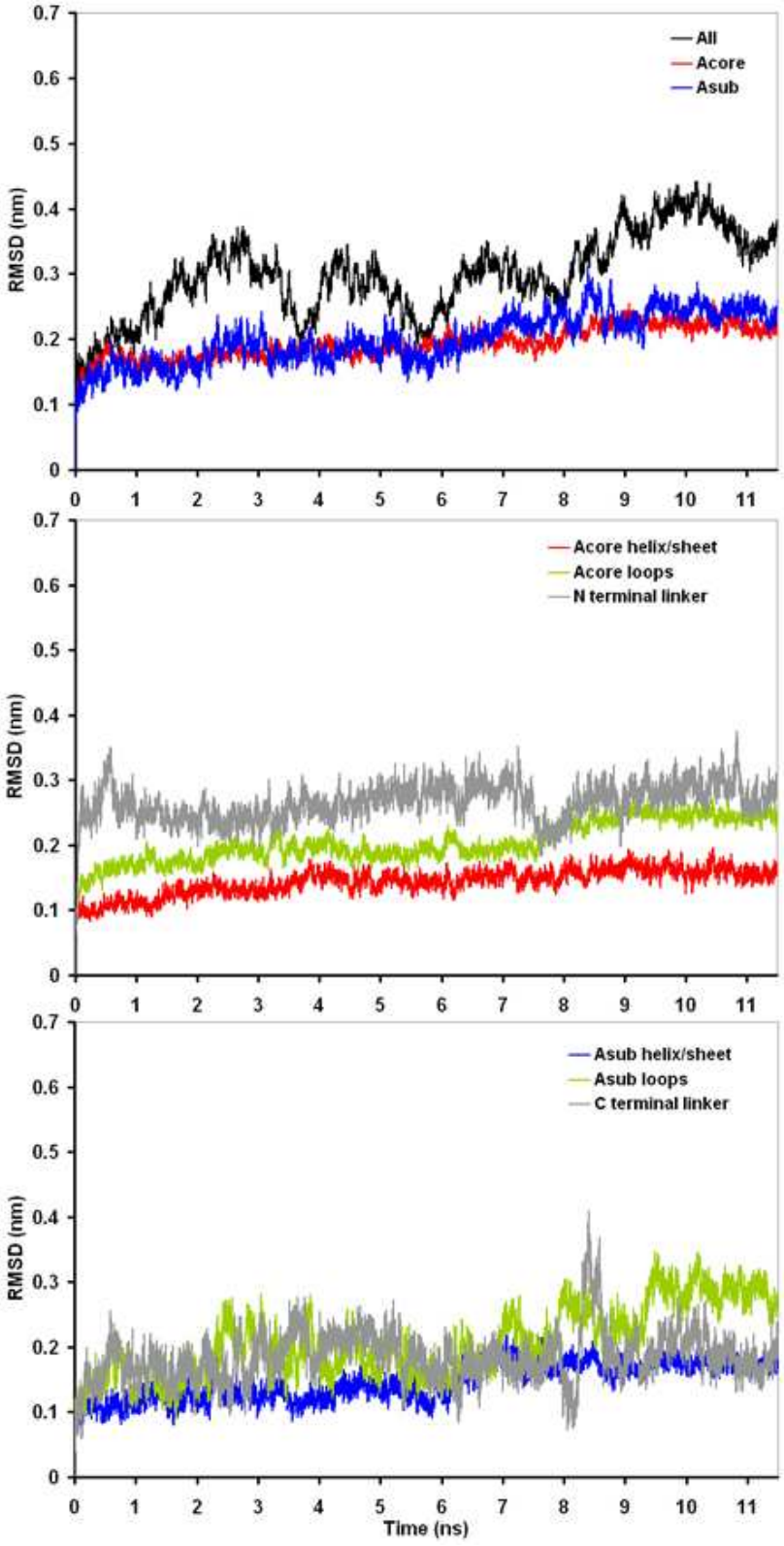


Figure 4.2: **RMSD PheA-Tyr simulation.** The conformational drift of PheA-Tyr, measured as C-α atom root mean square deviation (RMSD) from the starting structure. RMSDs vs. time are shown for the entire protein (black), the A_{core} domain (red) and A_{sub} domain (blue), in the upper graph. The RMSD of the decomposed linker (grey), secondary structural elements - helices and sheets (red), and loop regions (green) of the A_{core} domain in shown in the middle graph, and the RMSD of the equivalent regions of the A_{sub} domain in the lower graph.

the PheA-Tyr simulation than in either of the PheA holo cognate substrate simulations. As expected, greater structural drift is observed in the flexible loops, with the C-terminal linker region exhibiting the greatest drift from the starting structure.

PheA-Asp

Figure 4.3 shows the Cα atom RMSDs from the PheA-Asp simulation. The all Cα atom RMSD indicates little relative motion between the domains in this simulation.

The A_{core} domain is stable during the simulation. Greater structural drift is observed in the A_{sub} domain. The RMSD of this region is very similar to that of the full protein RMSD until the eighth nanosecond, when the RMSD of the A_{sub} domain begins to climb from $\sim$0.25 nm ending the simulation still increasing at $\sim$0.45 nm. This and the RMSD of the sheets and helices from the A_{sub} domain, which increases from 0.15 nm at 8 ns to 0.25 nm at 11.5 ns, indicates structural rearrangements of this domain in the final four nanoseconds of the simulation. The loops of the A_{sub} domain exhibit greater structural drift than observed in the PheA holo and PheA-Tyr simulations, particularly from the sixth nanosecond onwards.

The RMSDs of the A_{core} domain components are consistent with those observed in the PheA-Tyr and holo simulations.

PheA-Arg

The RMSD analysis from the PheA-Arg simulation is shown in figure 4.4.

As in the PheA-Asp simulation, the all Cα atom RMSD indicates little relative motion between the domains. During the first five nanoseconds of the simulation the RMSDs of the whole protein, A_{core} and A_{sub} domains are virtually identical. During the sixth nanosecond, the RMSD of both the all Cα atom and A_{sub} domain Cα atoms begins to increase, while that of the A_{core} domain remains constant. This trend suggests the changes in the overall RMSD can be attributed to changes in the A_{sub} domain. As in the PheA-Asp simulation, the RMSD of the A_{sub} domain is greater than that of the RMSD of the full protein in the

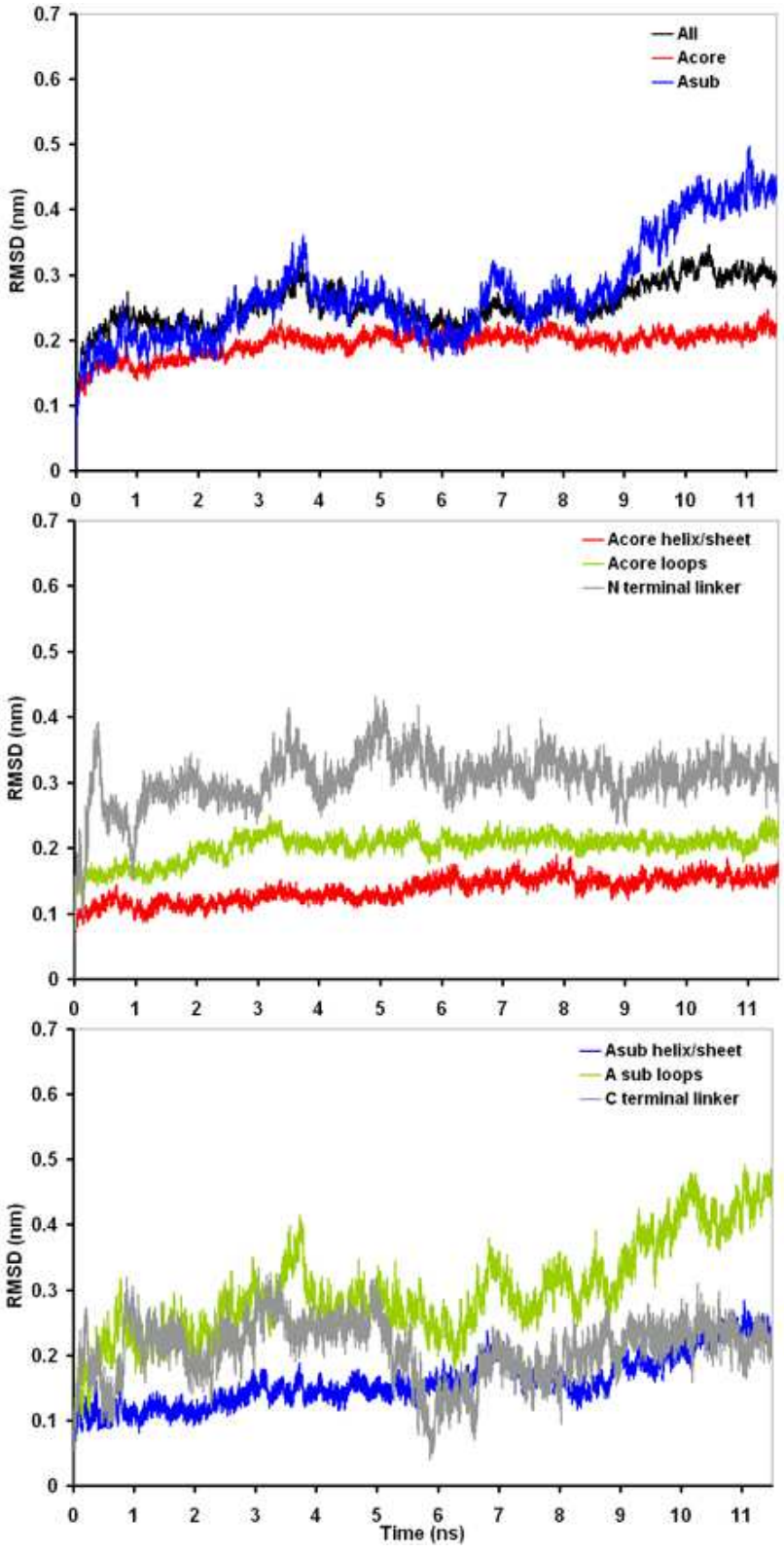


Figure 4.3: **RMSD PheA-Asp simulation.** The conformational drift of PheA-Asp, measured as C-α atom root mean square deviation (RMSD) from the starting structure. RMSDs vs. time are shown for the entire protein (black), the A_{core} domain (red) and A_{sub} domain (blue), in the upper graph. The RMSD of the linker (grey), secondary structural elements - helices and sheets (red), and loop regions (green) for the A_{core} domain and the A_{sub} are shown in the middle and lower graphs respectively.

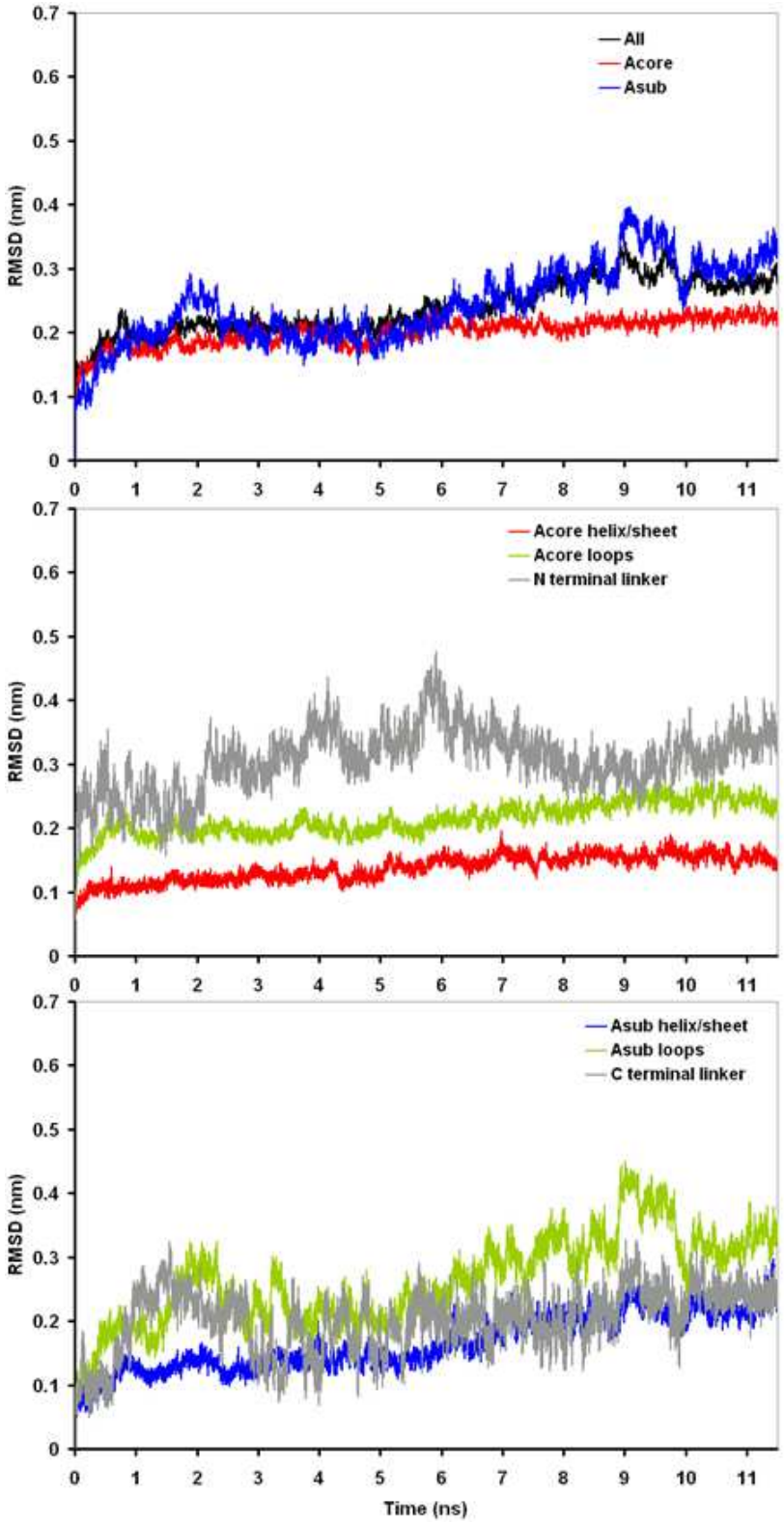


Figure 4.4: **RMSD PheA-Arg simulation.** The conformational drift of PheA-Arg, measured as Cα atom root mean square deviation (RMSD) from the starting structure. RMSDs vs. time are shown for the entire protein (black), the A_{core} domain (red) and A_{sub} domain (blue), in the upper graph. The RMSD of the decomposed linker (grey), secondary structural elements - helices and sheets (red), and loop regions (green) of the A_{core} domain in shown in the middle graph, and the RMSD of the equivalent regions of the A_{sub} domain in the lower graph.

final stages of the simulation.

The RMSDs of the component parts of the A_{core} domain are consistent with those observed in the other ligated PheA simulations. The RMSDs of the component parts of the A_{sub} domain are largely consistent with those from the PheA-Asp simulation. However less structural drift is observed in the A_{sub} domain loops in the PheA-Arg simulation than the PheA-Asp simulation.

4.4.3 Radius of Gyration

	Starting	PheA1-holo	PheA2-holo	PheA-Tyr	PheA-Asp	PheA-Arg
Rg (nm)	2.38	2.35 (0.02)	2.36 (0.03)	2.35 (0.02)	2.35 (0.01)	2.37 (0.01)

Table 4.4: **Average values of the radius of gyration** (Rg) for the PheA noncognate substrate simulations. Standard deviation in parentheses.

The average radius of gyration (Rg) was calculated for PheA in each simulation, table 4.4.3. The Rg, alongside the values obtained for PheA in the PheA-holo simulations, indicates that the protein retains overall compactness during the timescale of the simulations. Analysis of the radius of gyration throughout the simulation indicated the stability of the overall fold of PheA in each of the simulations; data included on the accompanying CD.

4.4.4 Secondary Structure

The secondary structure content of PheA, according to DSSP classification, from each of the simulations was calculated versus time; visual plots are included the accompanying CD. On the timescale of the simulation the core structure of PheA in each simulation is stable.

In the PheA apo and holo simulations presented in Chapter 3 variation in the secondary structure content of residues from the interdomain hinge region (motif A8) and A10 motif K loop was observed. This variation is also observed in the noncogante substrate simulations.

Secondary Structure - A8 motif residues

In the PheA holo simulations residues 407–417 at the interdomain region, identified in the crystal structure to form two short β-strands, merge to form a single long β-strand throughout the simulation. This β-strand is observed in PheA ligated with the noncognate substrates, however, it is not maintained for as long.

In the PheA-Tyr simulation residues 407–417 form the long β-strand between 1 and 4 ns, thereafter it is observed intermittently. The strand is formed for a longer duration in the PheA-Asp simulation; it is present for the first nine nanoseconds, absent during the tenth nanosecond and thereafter formed intermittently. In the PheA-Arg simulation these residues form two distinct strands between 0 and 4.6 ns and thereafter the single / two strand structure is adopted intermittently.

Secondary Structure - A10 motif residues

The β sheet that is formed by the residues either side of the A10 motif K loop in the PheA1-holo, and PheA apo simulations, and briefly in the PheA2-holo simulation, is observed in each of the noncognate substrate simulations for varying lengths of time. It is observed for the longest in the PheA-Tyr simulation, where it is formed between 3.2 and 8.8 ns, in the PheA-Asp simulation it is formed between 9.25 and 10.6 ns and in the PheA-Arg simulation it is formed between 6.45 and 7.2 ns.

Secondary Structure Residues on the Exterior of the Protein

Fluctuating secondary structure was observed in the apo and holo simulations for a number of residues located on the exterior of the protein, far from the interdomain hinge, domain interface and active site. Variation in the secondary structure of these regions is also observed in the cognate ligand simulations although to a lesser extent.

Helix H5, formed by residues 374 to 380, fluctuates between α and π helical structure and turn/alpha-helix/turn structure in the PheA2-holo simulation. This helix is stable in the

PheA-Tyr and PheA-Asp simulations. In the PheA-Arg simulation these residues predominantly form a πhelix structure from 4.4 ns until the end of the simulation.

The long unstructured region of sequence linking β strands A4 to A5, shown to form a short α-helix in the PheA2-apo, PheA1-holo and PheA2-holo simulations but not in the PheA1-apo simulation, predominantly forms an α helix in the PheA-Tyr and PheA-Arg simulations. In the PheA-Asp simulation an αhelix is formed by this region from 5.3 ns until the end of the simulation.

4.4.5 Structural Flexibility

The RMSFs of the C-α atoms of the simulated structures relative to the average structure were calculated for PheA in each simulation to provide an indication of the relative flexibility of the different regions of the protein. The RMSFs of PheA in each noncogante simulation are shown overlayed with those from the PheA holo simulations in figure 3.6 to provide a comparison of the relative flexibility of PheA.

Regions of notably different flexibility from the PheA-holo simulations, include the A3 motif loop, the binding pocket residues and the residues of the A_{sub} domain.

The RMSFs of the A_{sub} domain residues in the PheA-Asp and PheA-Arg simulations show this region is less flexible than observed in either of the PheA-holo simulations presented in Chapter 3. This also provides an indication that the domain motion observed in the PheA simulation with the cognate substrate may not occur in the PheA-Arg and PheA-Asp simulations. The flexibility of the A_{sub} domain residues in the PheA-Tyr simulation is similar to that observed in the PheA2-holo simulation.

The A3 motif loop in the PheA-Tyr simulation exhibits greater flexibility than in either holo simulation; 0.5 nm in PheA-Tyr, as compared to 0.3 nm in PheA1-holo and 0.2 nm in PheA2-holo. The A3 motif loop is also flexible in the Phe-Asp simulation, 0.35 nm however this region in the PheA-Arg simulation is of comparative flexibility to that observed in the PheA1-holo simulation.

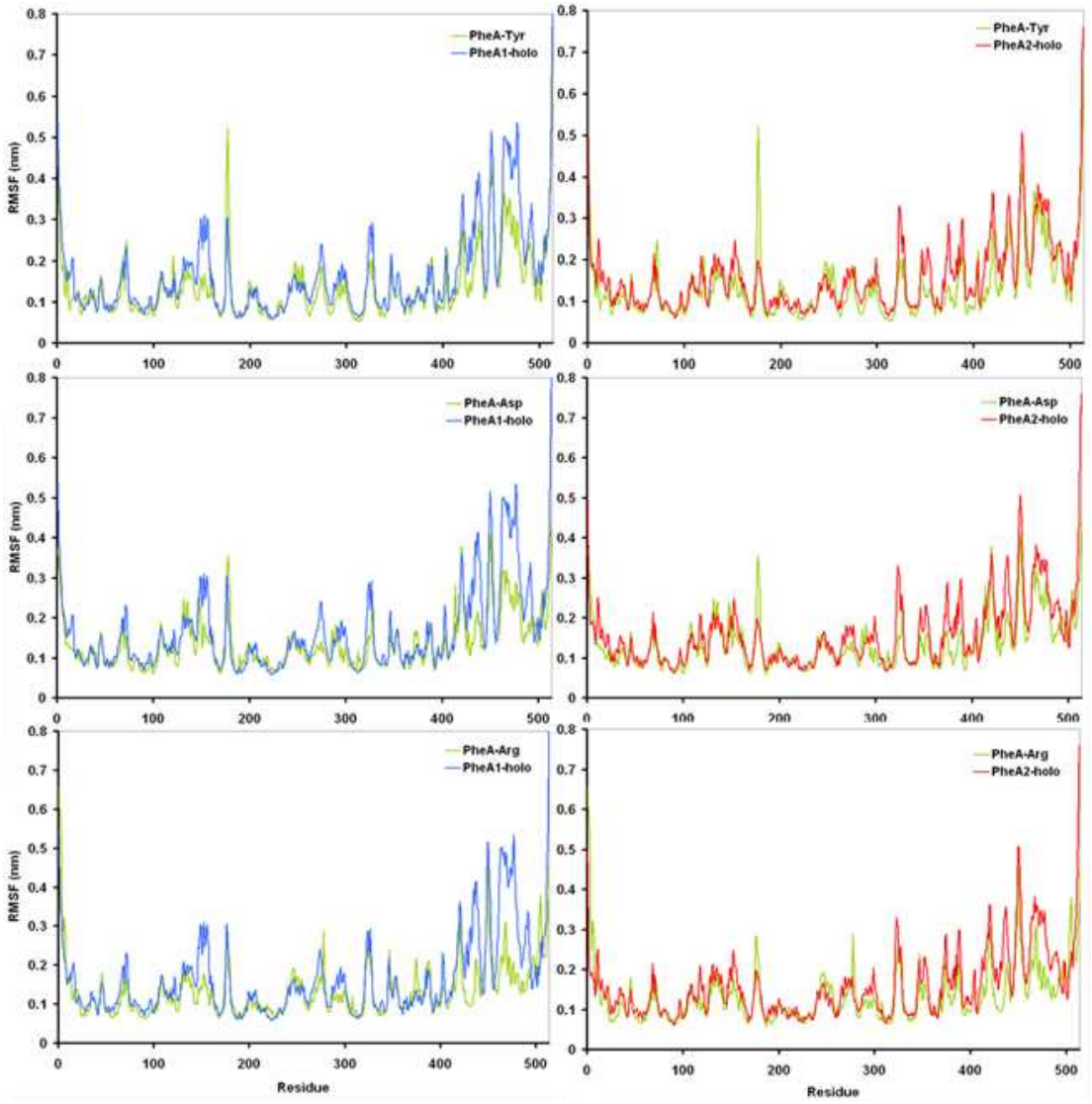


Figure 4.5: **RMSFs of the PheA-Tyr, PheA-Asp and PheA-Arg simulations.** The time-averaged Cα RMSFs as a function of residue number for the PheA-Tyr (upper), PheA-Asp (middle), and PheA-Arg (lower) simulations shown overlayed with the RMSFs of PheA from the PheA1-holo and PheA2-holo simulations.

The key binding pocket residues (located in the region 219–315) exhibit greater flexibility in the Tyr simulation and the PheA-Arg simulation than in the PheA-holo simulations. This flexibility is not observed in the PheA-Asp simulation.

4.4.6 Principal Modes of Motion

As for the cognate holo simulations, the principal modes of motion of PheA for each noncognate substrate simulation were identified using principal components analysis. The PCA was performed by least squares fitting to the backbone atoms.

Figure 4.6 describes the size of each of the ten first eigenvectors (index). The first eigenvector from each simulation is less than half the size of those from the PheA-holo simulations.

The eigenvectors were projected onto the trajectory and the structures equating to the extremes of motion in each eigenvector obtained. These structures were analysed using the DynDom server[274,275] to identify whether the principal modes of motion were between the domains of the protein. The orientation of PheA used to define the left and right hand sides of PheA is shown in figure 3.14 of Chapter 3.

Principal Modes of Motion of PheA-Tyr

Figure 4.7 shows the structures corresponding to the extremes of motion for the first three eigenvectors in the PheA-Tyr simulation. The first eigenvector describes the rotation and tipping of the A_{sub} domain towards the A3 motif loop and the left of the A_{core} domain about hinge residues 413 to 414 (A8 motif) and 495 to 497 (residues preceding the A10 motif). The A_{sub} domain rotates clockwise by 42° about the axis defined by the hinge residues. The direction of this motion is consistent with that described by the principal eigenvector from the PheA-holo simulations, i.e. towards the A3 motif loop. The domain division of PheA is similar to that observed in the second and first eigenvector of PheA1-holo and PheA2-holo respectively. The extremes of this motion are seen at 0.217 and 10.188 ns.

The second eigenvector of PheA-Tyr describes the tipping of the A_{sub} domain towards the

Index	PheA-Tyr Eigenvalue (nm^2)	PheA-Tyr Cumulative (%)	PheA-Arg Eigenvalue (nm^2)	PheA-Arg Cumulative (%)	PheA-Asp Eigenvalue (nm^2)	PheA-Asp Cumulative (%)
1	12.40	34.25	13.31	38.39	9.73	29.00
2	4.53	46.75	3.31	47.92	5.16	44.38
3	2.44	53.50	2.06	53.86	3.29	54.18
4	1.76	58.35	1.84	59.16	1.85	59.68
5	1.54	62.60	1.23	62.70	1.08	62.90
6	1.07	65.56	0.96	65.47	0.76	65.16
7	0.93	68.13	0.71	67.52	0.66	67.11
8	0.66	69.94	0.60	69.25	0.54	68.71
9	0.55	71.47	0.54	70.79	0.50	70.19
10	0.47	72.76	0.43	72.03	0.44	71.50

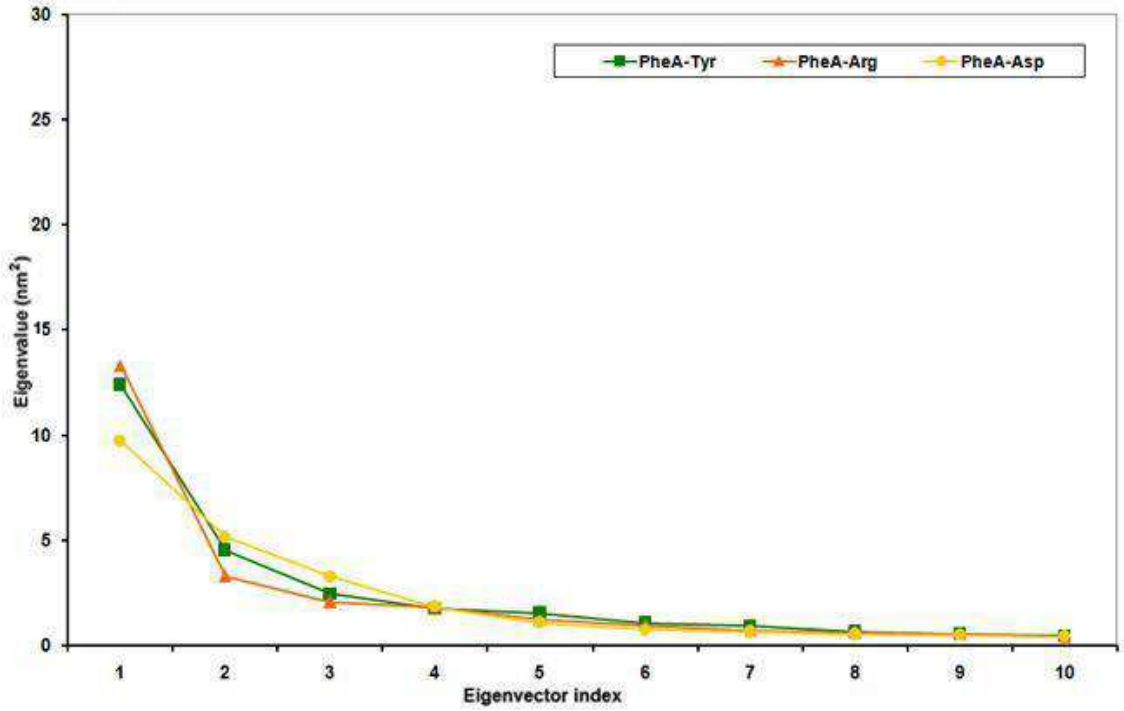


Figure 4.6: **PCA analysis of the PheA-Tyr, -Asp and -Arg simulations**The eigenvectors (index) and eigenvalues of the PheA-Tyr (green), PheA-Asp (yellow), and PheA-Arg (orange) simulations.

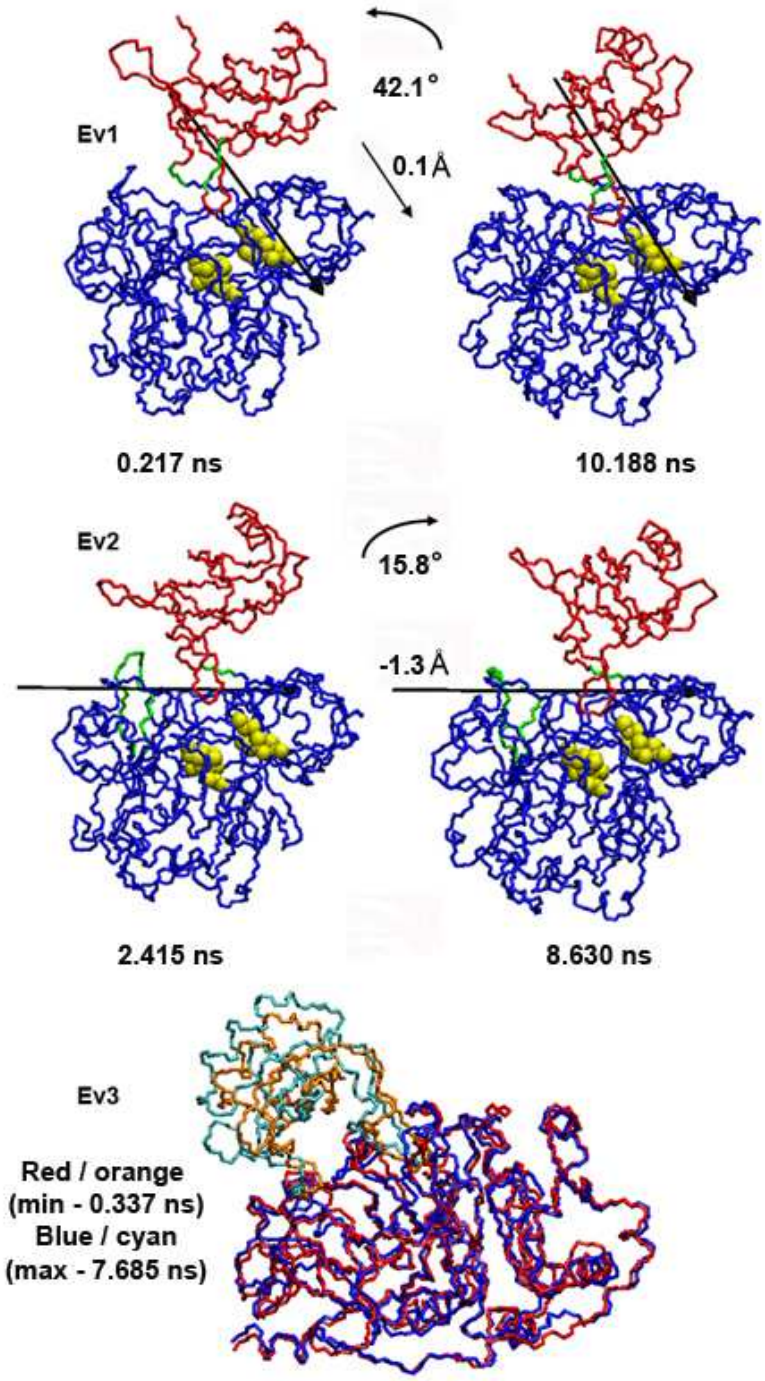


Figure 4.7: **Domain motion in the PheA-Tyr simulation.** Interdomain motion in the PheA-Tyr simulation; corresponding to the first three eigenvectors. The motion in the first two was identified and described by DynDom. Domain 1 (static) is shown in blue, domain 2 (moving) in red, the hinge regions in green and the phenylalanine binding pocket residues in yellow using VDW representation. The movement identified by eigenvector 3 (Ev 3) is shown by overlaying the structures of the extreme projections of this eigenvector.

right side of PheA, away from the A3 motif loop. The extremes of this motion are observed at 2.415 and 8.630 ns. The hinge resides for this motion are from the A8 motif (411–412) at the interdomain boundary and from the A3 motif loop (172–184). The direction of this motion is similar to that observed in the PheA1-holo simulation and in each of these simulations, PheA-Tyr and PheA1-holo, the A3 motif loop is highly flexible.

Dyndom analysis of the extreme projection of the trajectory along the third eigenvectors of the PheA-Tyr simulation did not determine any interdomain motion. Overlaying of the structures of the extreme projections of this eigenvector and fitting of the A_{core} domain to self suggests that this eigenvector may describe a slight backwards tilting of the A_{sub} domain towards the A8 motif interdomain linker residues.

Principal Modes of Motion of PheA-Asp

The extremes of motion for the first three eigenvectors are observed at; 0.793 and 11.046 ns, 7.267 and 10.396 ns, and 0.851 and 4.310 ns. DynDom only identified domain motion between the structures equating to the motion described by the second eigenvector. This motion was indentified between three domains that do not correlate with any of the PheA subdomains; see figure 4.8.

Overlay of the extreme structures from the first eigenvector indicated some rearrangements in the A_{sub} domain including a shift in the positioning of the A10 K loop and A3 motif loop. This corresponds with a small peak observed in both the all C-alpha Asub domain RMSD and the Asub domain loops RMSD at 11 ns.

Principal Modes of Motion of PheA-Arg

DynDom analysis of the extreme projections of the trajectory along the first three eigenvectors of the PheA-Arg simulation did not reveal any interdomain motion. The extremes of motion for these eigenvectors were observed at; 0.924 and 10.204 ns, 7.059 and 9.665 ns and 0.326 and 2.997 ns.

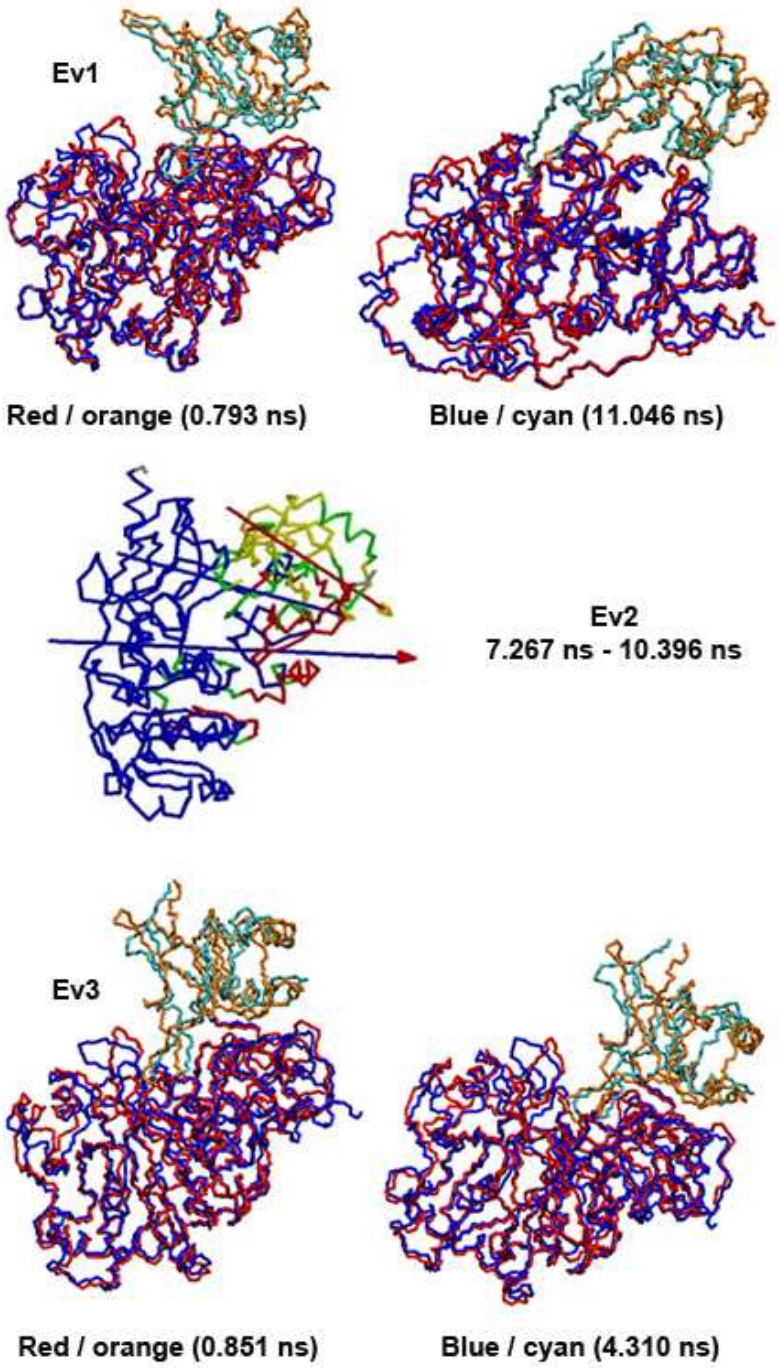


Figure 4.8: **Domain motion in the PheA-Asp simulation.** Interdomain motion in the PheA-Asp simulation; corresponding to the first three eigenvectors. The motion of eigenvector 2 was identified and described by DynDom. Here domain 1 (static) is shown in blue, domain 2 (moving) in red, domain 3 (moving) in yellow, and the hinge regions in green. The movement identified by eigenvectors 1 (Ev 1) and 3 (Ev 3) is shown by overlaying the structures of the extreme projections of these eigenvector. These overlays are shown from differing angles.

Visualisation, by overlaying the structures of the first three eigenvectors of the PheA-Arg simulation and fitting to the A_{core} domain, did not identify any obvious structural changes associated with these principal components.

4.4.7 Intramolecular Hydrogen Bonding

The average number of intramoleclar hydrogen bonds formed between PheA during the entire simulation, first nanosecond and last nanosecond in each simulation are presented in table 4.4.7. The number of hydrogen bonds formed between PheA in each system increases slightly throughout the simulation. A plot of the intramolecular hydrogen bonding versus time shows the number of hydrogen bonds increases steadily throughout each simulation, data on accompanying CD.

	PheA-Tyr	PheA-Asp	PheA-Arg
Whole (SD)	726.10 (16.1)	732.83 (16.6)	736.40 (17.8)
1st ns (SD)	718.2 (16.4)	725.3 (15.5)	723.4 (15.6)
Last ns (SD)	736.6 (15.6)	736 (14)	739.5 (15.4)

Table 4.5: Intramolecular (protein-protein) hydrogen bonds for the PheA noncognate substrate simulations. Standard deviation in parentheses

4.4.8 Interdomain Hydrogen Bonding

Hydrogen bonding between residues of the A_{core} and A_{sub} domain was assessed in each of the noncognate ligand simulations, see figure 4.9. Definition of the regions of interdomain hydrogen bonding is provided in section 3.3.10 of chapter 3.

In contrast with the PheA-holo simulations very few interdomain hydrogen bonding interactions are observed between the A3 motif or residues on the left of the PheA in any of the noncognate ligand simulations. This is particularly interesting in the PheA-Tyr simulation where the first eigenvector described motion of the A_{sub} domain towards the A3 motif loop. In the PheA-Tyr simulation the vast majority of interdomain hydrogen bonding interactions are observed between residues in the interdomain hinge region. Residues involved in these

interactions include Arg 412, Asp 414 and Glu 416. The number of hydrogen bonding interactions formed in this region peaks between the time that the extremes of motion described by the second eigenvector are observed.

Interdomain hydrogen bonding in the PheA-Asp and PheA-Arg simulations occurs between the hinge region, A10 motif K loop and residues located on the right side of PheA/A3 motif loop.

4.4.9 Substrate Hydrogen Bonding

Hydrogen bonding interactions between PheA and the noncognate substrates, and the key Asp 219 (pdb: 235) and Lys 501 (pdb: 517) binding pocket residues and PheA were assessed in each simulation to provide a measure of the substrate binding. The average number of hydrogen bonds present per nanosecond has been used as a measure of the strength of the hydrogen bonding interactions between particular residue groups.

PheA-Tyr

The hydrogen bonding interactions between Tyr and PheA, figure 4.10, are similar to those observed for the Phe substrate in the PheA1-holo simulation. The hydrogen bonding of the L-Tyr αamino group with the carboxyl sidechain of Asp 219 is, however, stronger than that observed in PheA1-holo and conversely the interaction of the L-Tyr αcarboxyl group with the Lys 501 amino sidechain group is weaker.

Hydrogen bonding patterns between the Asp 219 residue and PheA are similar to those observed in the PheA1-holo simulation. The Asp 219 residue is stablised by hydrogen bonding interactions of strength 0.6–0.8 with the main chain amino group of Ala 220, and a weaker hydrogen bonding interaction (0.3–0.45) with the main chain amino group of Asp 219.

In contrast with what is seen in both PheA1- and PheA2-holo, stronger hydrogen bonding is observed between Lys 501 amino sidechain group and the main chain carbonyl group

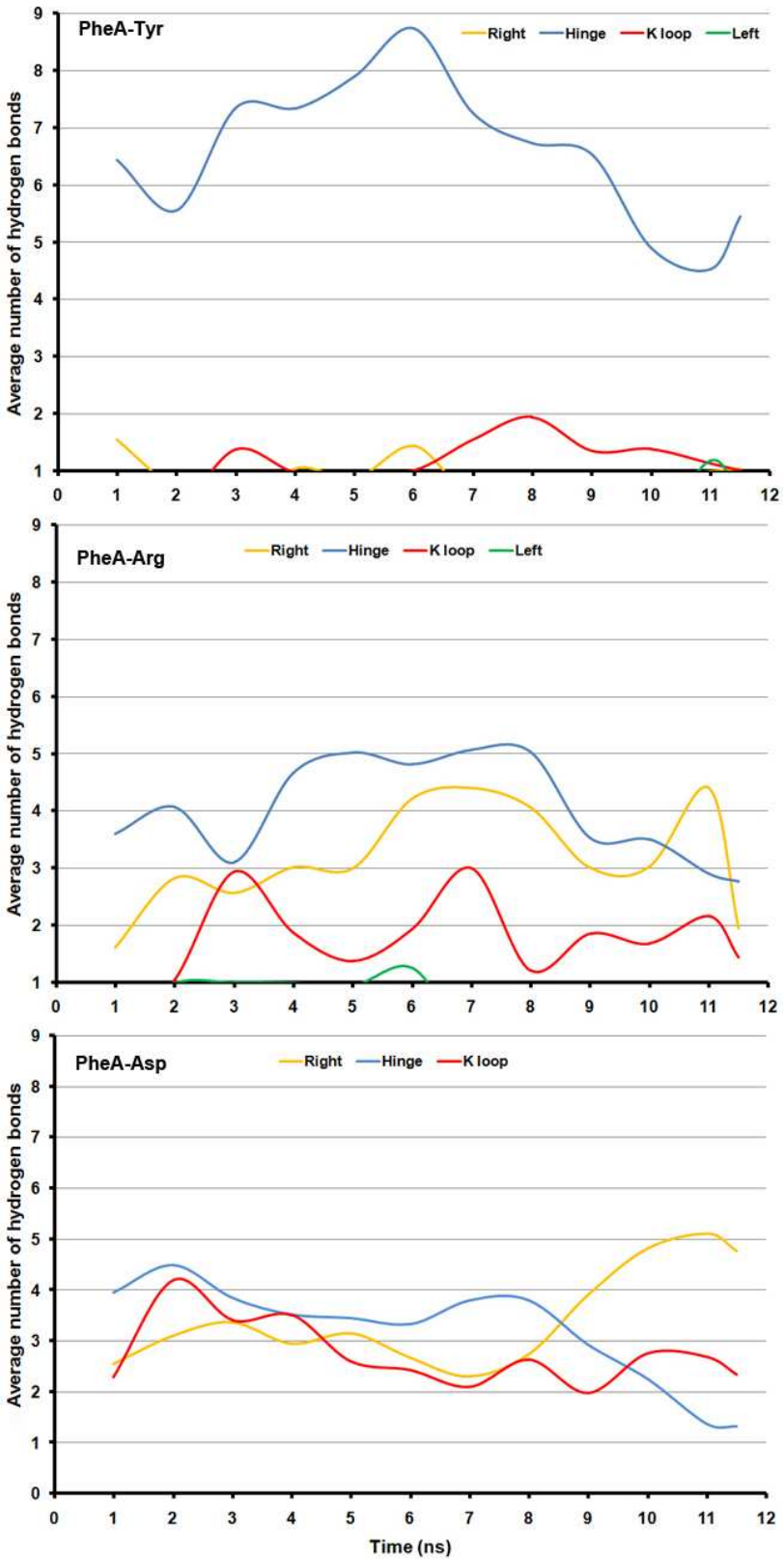


Figure 4.9: **Interdomain hydrogen bonding in the PheA-Tyr -Arg and -Asp simulations.** The average number of interdomain hydrogen bonds between the specified regions of the PheA structure on a nanosecond timescale in the PheA-Tyr (upper), PheA-Arg (middle), and PheA-Asp (lower) simulations. Hydrogen bonds formed at the hinge / interdomain linker region are shown in blue, those formed on the left side of PheA in green, those on the right side in orange, and those between one residue of the K loop and the A_{core} domain of PheA in red.

of Gly 286. The strength of this interaction increases on the time scale of the simulation, starting at 0.1 and ending the simulation at 0.5.

No hydrogen bonding interactions between the L-Tyr substrate hydroxyl sidechain and PheA are observed during the simulation.

PheA-Asp

Analysis of the substrate-PheA binding in the PheA-Asp simulation, is shown in figure 4.11. The hydrogen bonding interaction of the L-Asp substrate with the Asp 219 residue is lost during the fourth nanosecond of the simulation. This correlates with the time the extreme motion of described by the third eigenvector is observed.

Hydrogen bonding between the L-Asp substrate α-carboxyl group with PheA is disordered throughout the simulation. No hydrogen bonding is observed between this group and the invarient Lys 501 residue. During the first three nanoseconds of the simulation strong hydrogen bonding is observed between the L-Asp substrate α-carboxy group and Thr 310, with interactions formed with the amino main chain atom and hydroxyl sidechain. These interactions are lost during the fourth nanosecond which is when the extreme motion described by the third eigenvector is observed. A number of interactions of varying strength and persistence are formed between the L-Asp substrate carboxyl group and residues from the A3 loop motif, from the fourth nanosecond onwards.

The hydrogen bonding interactions formed between the Asp 219 α-carboxyl group and Asp 219 amino main chain group, and Ala 220 main chain amino group of PheA in the PheA-Asp simulation are similar in strength and persistence to those observed in the PheA2-holo simulation.

During the first two nanoseconds of the simulation a strong hydrogen bonding interaction is observed between the sidechain group of the L-Asp substrate and the amino group of Lys 501. This interaction weakens during the third nanosecond, is absent by the fourth nanosecond, increasing to a strength of $\sim$1 in the fifth nanosecond, after which time no interaction is observed between these two groups.

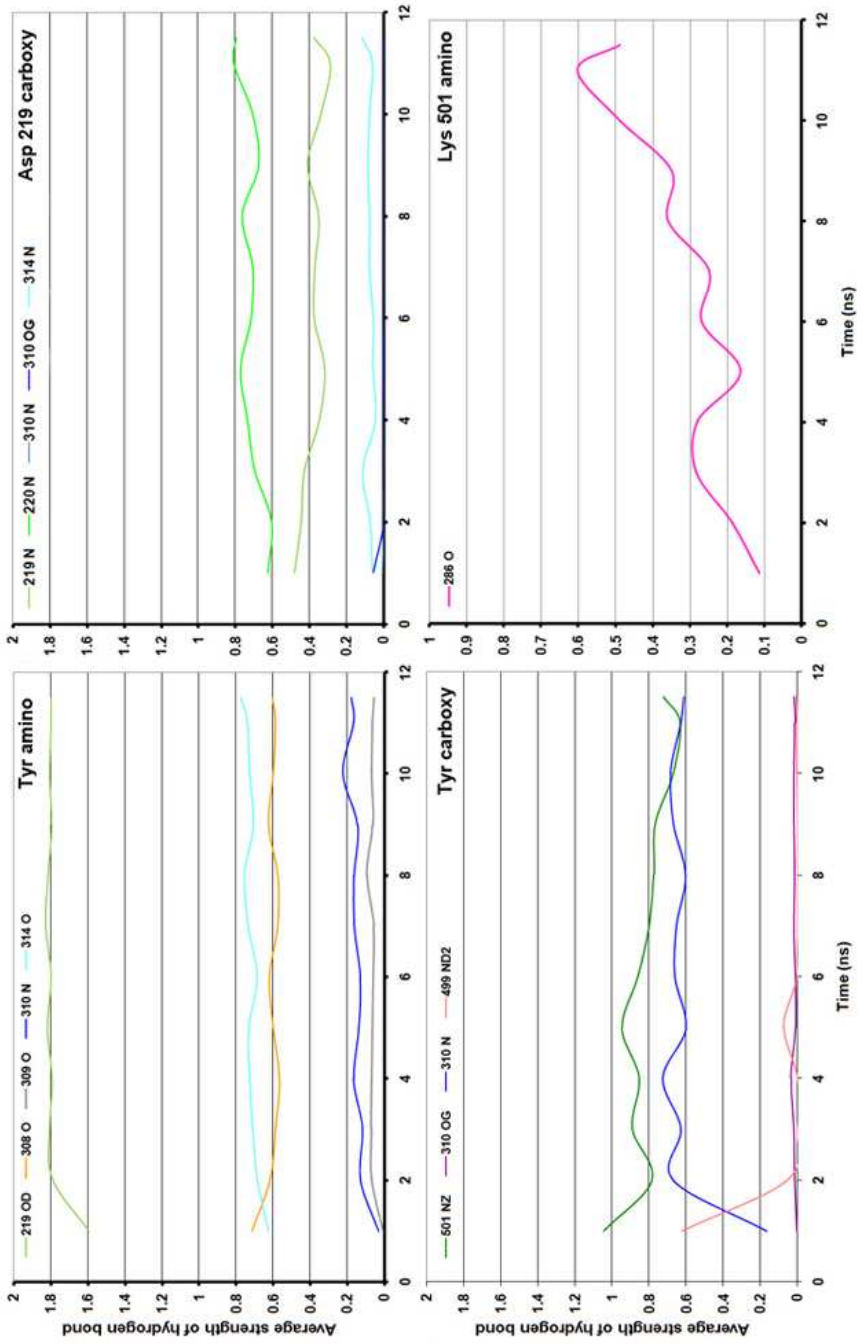


Figure 4.10: **Hydrogen bonding between the L-Tyr substrate and PheA in the PheA-Tyr simulation.** The upper left graph shows the average strength of the hydrogen bond(s) formed between the Tyr substrate amino group and the specified groups of PheA. The lower left graph shows the average strength of the hydrogen bond(s) formed between the Tyr substrate carboxy group and the specified groups of PheA. The upper right graph shows the average strength of the hydrogen bond(s) formed between the Asp 219 PheA carboxy group and the specified groups of PheA. The lower right graph shows the average strength of the hydrogen bond(s) formed between the Lys 501 PheA amino group and the specified groups of PheA.

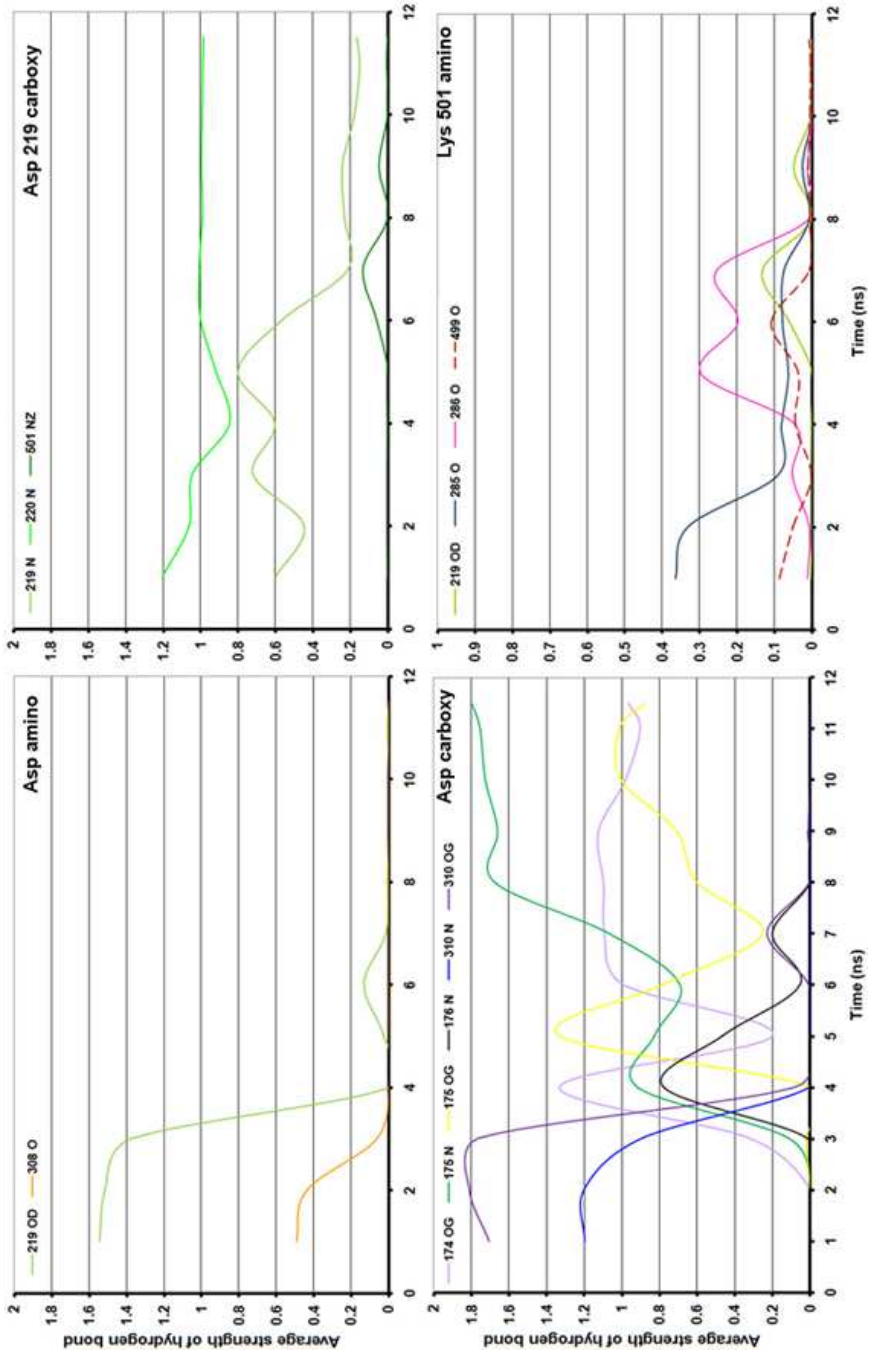


Figure 4.11: **Hydrogen bonding between the L-Asp substrate and PheA in the PheA-Asp simulation.** The upper left graph shows the average strength of the hydrogen bond(s) formed between the Asp ligand amino group and the specified groups of PheA. The lower left graph shows the average strength of the hydrogen bond(s) formed between the Asp ligand carboxy group and the specified groups of PheA. The upper right graph shows the average strength of the hydrogen bond(s) formed between the Asp 219 PheA carboxy group and the specified groups of PheA. The lower right graph shows the average strength of the hydrogen bond(s) formed between the Lys 501 PheA amino group and the specified groups of PheA.

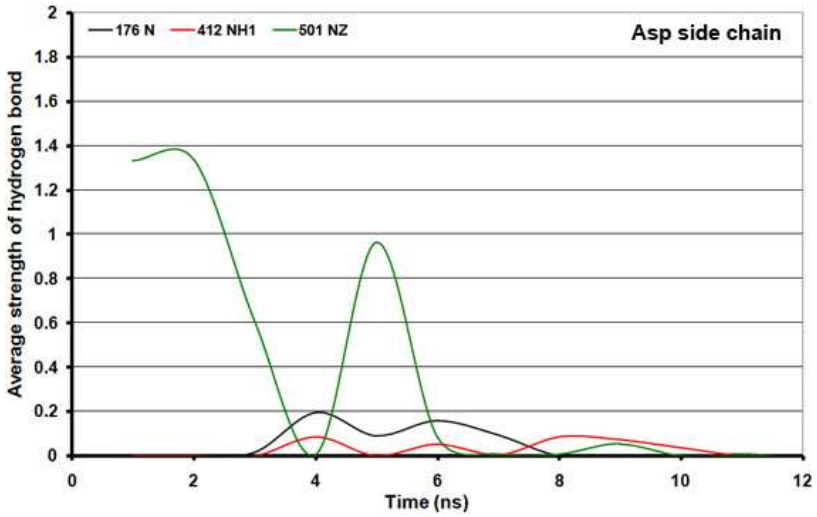


Figure 4.12: **Hydrogen bonding between the L-Asp substrate sidechain and PheA in the PheA-Asp simulation.**

To further understand the hydrogen bonding interactions observed between PheA and the L-Asp substrate, the PheA-Asp complex was visualised at different times throughout the simulation, see figure 4.13. These images show the substrate leaving the enzyme binding pocket and forming hydrogen bonding interactions with residues from the A3 motif loop. An overlay of PheA at the start and end of the simulation shows a shift in the positioning of the A10 motif K loop.

PheA-Arg

Figure 4.14 shows the hydrogen bonding interactions formed between the α-amino and α-carboxyl group of the L-Arg substrate and PheA on the time scale of the simulation.

The key hydrogen bonding interaction between the Arg substrate amino group and highly conserved Asp 219 of PheA is initially weak (0.1–0.4), however increases in strength from the fifth nanosecond of the simulation. After the eighth nanosecond the strength of this interaction fluctuates between 0.3 and 0.6. Initially one hydrogen bond is formed between the Arg substrate carboxy group and Lys 501 amino group. The strength of this hydrogen bond decreases over the time scale of the simulation to 0.

The hydrogen bonding interactions formed between the Asp 219 carboxyl group and Asp 219 amino main chain group of PheA in the PheA-Arg simulation is similar in strength

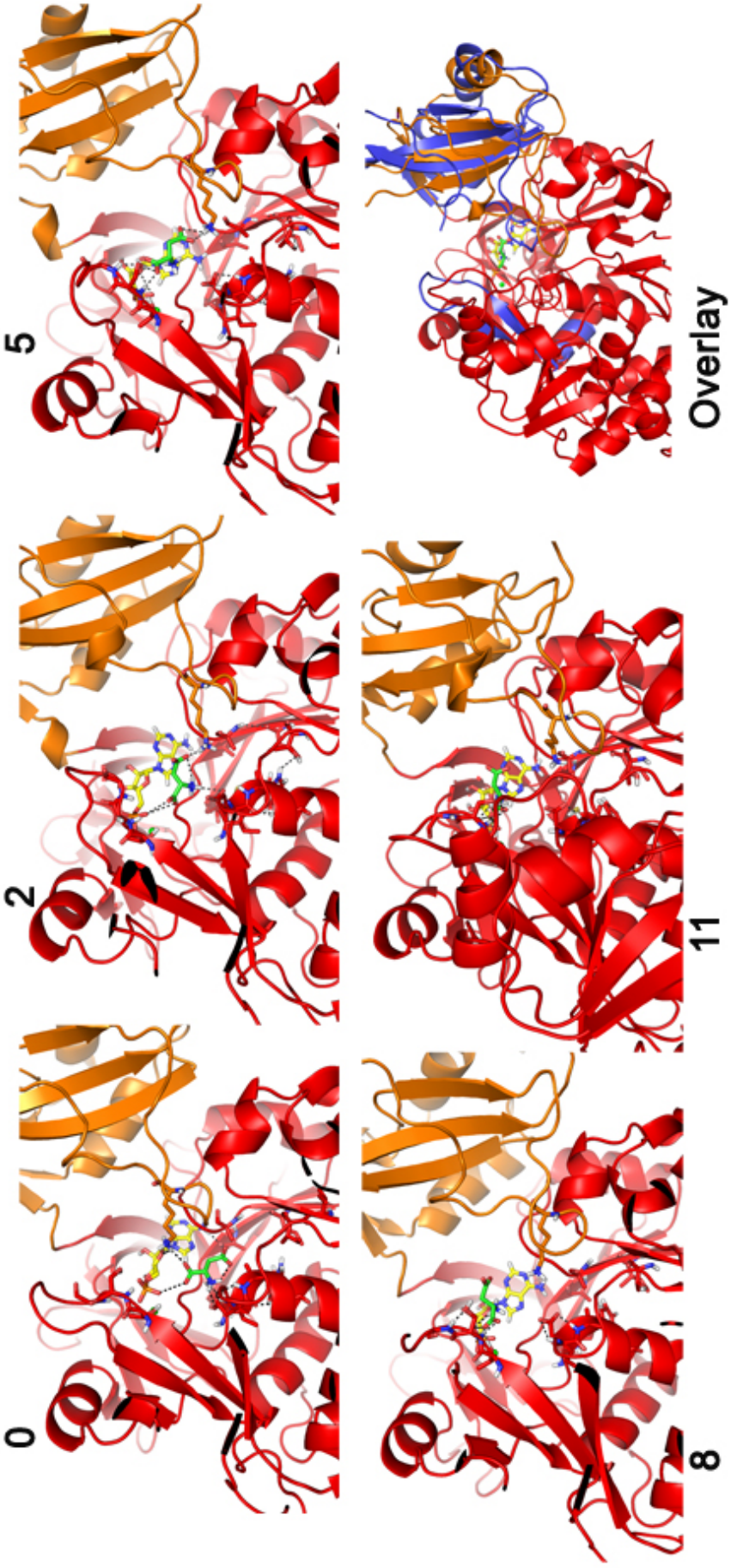


Figure 4.13: **Hydrogen bonding between the L-Asp substrate and PheA in the PheA-Asp simulation.** Shown at 0, 2, 5, 8 and 11 ns. The starting and ending structures from the simulation were overlayed to show the relative positioning of the A10 motif K loop.

and persistence to that observed in the PheA2-holo simulation. The hydrogen bonding interaction between the Asp 219 carboxyl group and Thr 174 hydroxyl sidechain group is formed later in the PheA-Arg simulation than in the PheA2-holo simulation; this interaction begins to form during the fifth nanosecond, by the eighth nanosecond one hydrogen bond has formed between these groups and this is maintained until the end of the simulation.

No hydrogen bonding interactions of notable duration are formed between the sidechain groups of the Arg substrate and PheA, see figure 4.15.

4.4.10 AMP Substrate Hydrogen Bonding

Adenine Binding

Analysis of the binding of the AMP ligand in each of the noncognate ligand simulations will be discussed with reference to the key interactions expected, both from findings in the literature[62] and those observed in the PheA holo simulations.

The hydrogen bonding interactions between the adenine moiety and PheA in the PheA-Tyr simulation, figure 4.16, are similar to those in the PheA holo simulations. Hydrogen bonding interactions between; the exocyclic nitrogen of AMP and main chain carbonyl of Ala 306, the N7 nitrogen of AMP and main chain amino of Gly 308, and the N7 nitrogen of AMP and main chain amino of Gly 286, are of a similar strength and duration to those observed in the PheA holo simulations. Overall fewer hydrogen bonding interactions are formed between the adenine moeity of AMP and PheA in the PheA-Tyr simulation than in the PheA-holo cognate simulations.

No stable hydrogen bonding interactions were observed between the adenine moiety of AMP and the PheA protein in the PheA-Asp simulation, see figure 4.17.

Some of the key hydrogen bonds between the adenine moiety of AMP and the PheA protein were observed in the PheA-Arg simulation, see figure 4.18. The interaction between the exocyclic N6 AMP atom (by which the protein selects for adenine over guanine[62]) and Ala

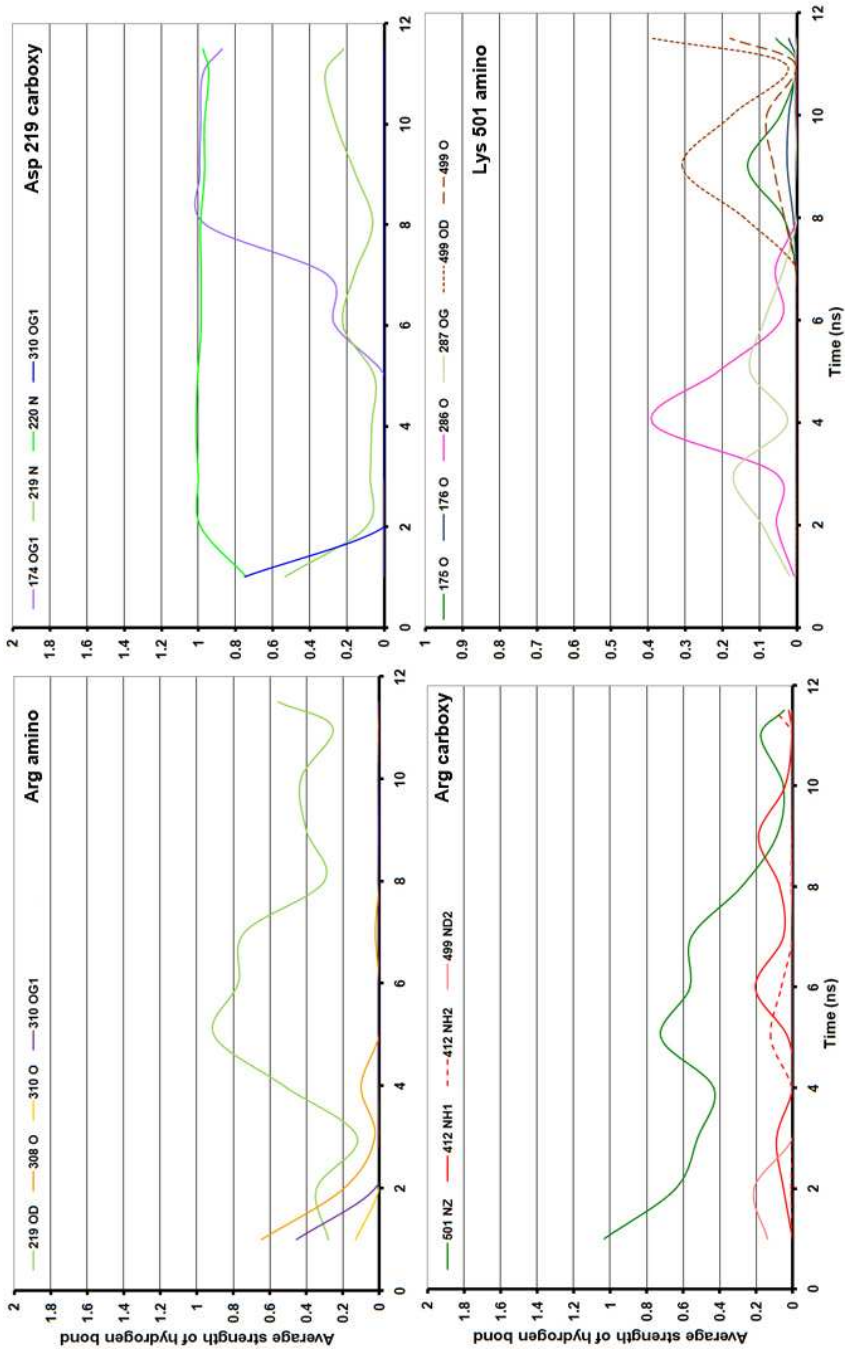


Figure 4.14: **Hydrogen bonding between the L-Arg substrate and PheA in the PheA-Arg simulation.** The upper left graph shows the average strength of the hydrogen bond(s) formed between the Arg ligand amino group and the specified groups of PheA. The lower left graph shows the average strength of the hydrogen bond(s) formed between the Arg ligand carboxy group and the specified groups of PheA. The upper right graph shows the average strength of the hydrogen bond(s) formed between the Asp 219 PheA carboxyl group and the specified groups of PheA. The lower right graph shows the average strength of the hydrogen bond(s) formed between the Lys 501 PheA amino group and the specified groups of PheA.

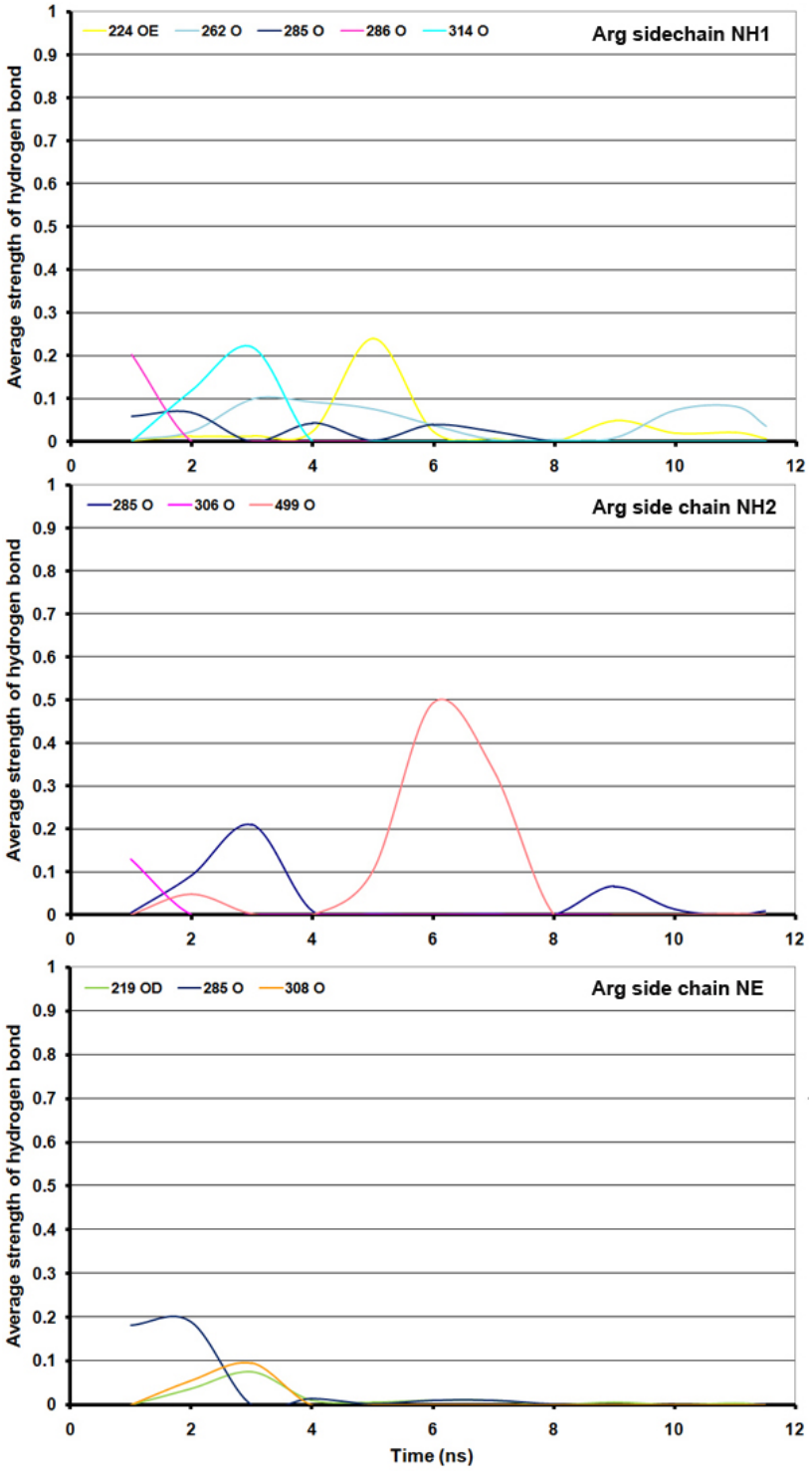


Figure 4.15: **Hydrogen bonding between the L-Arg substrate sidechain groups and PheA in the PheA-Arg simulation.**

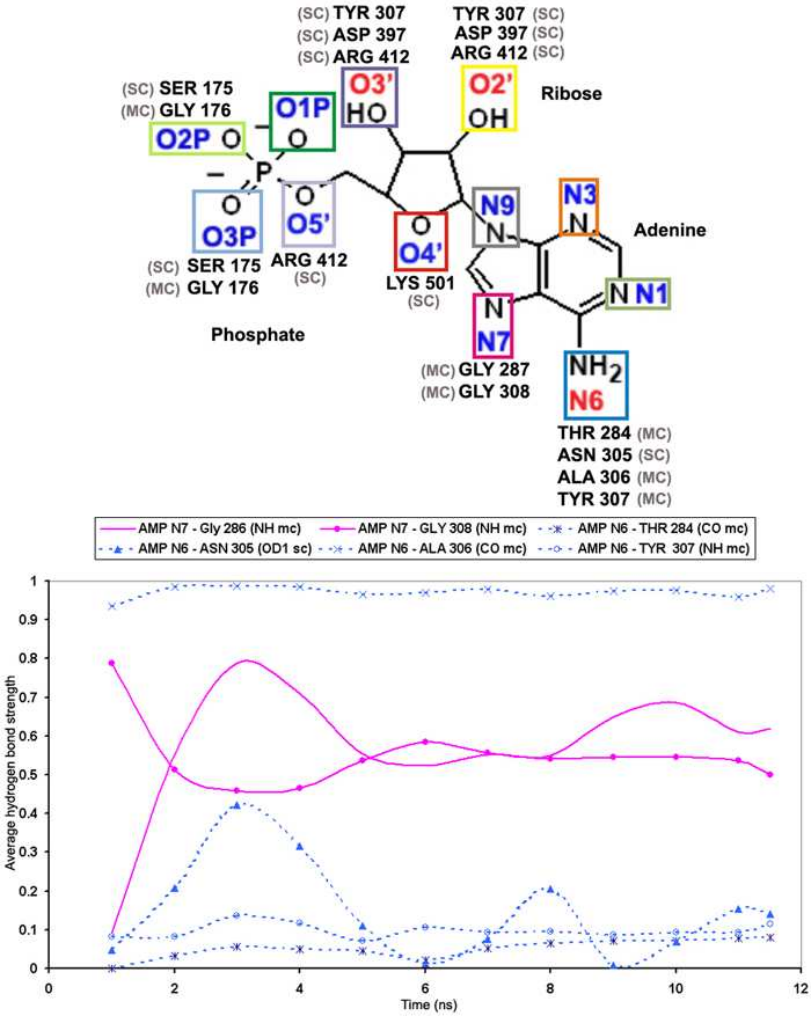


Figure 4.16: **Hydrogen bonding between the adenine moiety of the AMP ligand and PheA in the PheA-Tyr simulation. Dashed lines represent interactions where the AMP atom / group is a donor, and solid lines represent interactions where AMP is an acceptor.**

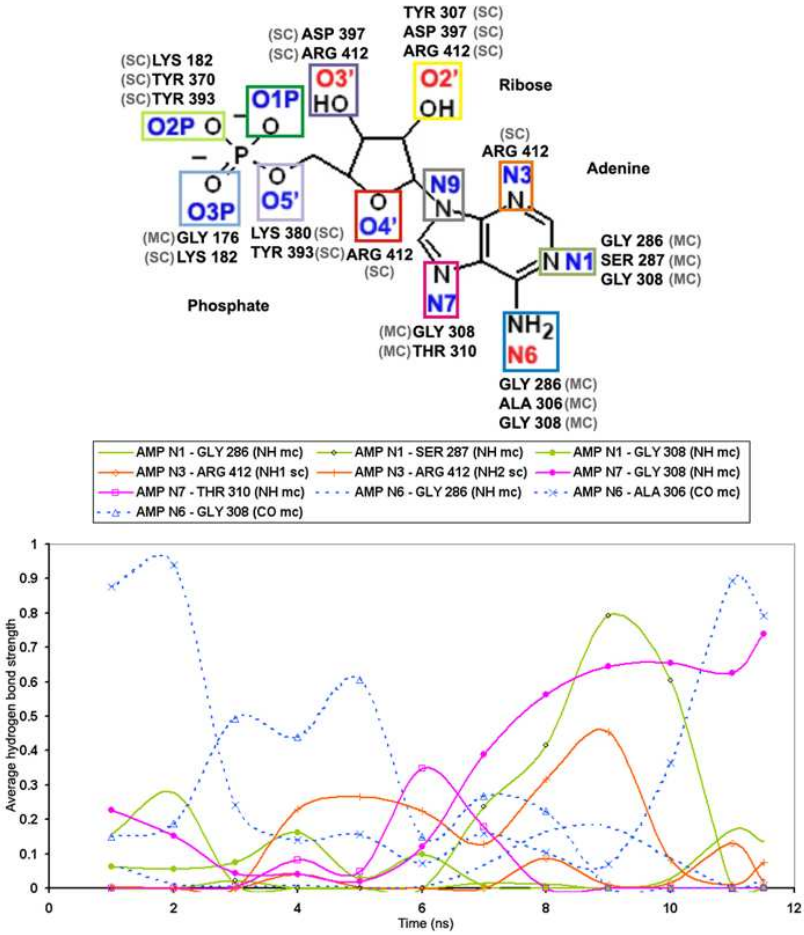


Figure 4.17: **Hydrogen bonding between the adenine moiety of the AMP ligand and PheA in the PheA-Asp simulation. Dashed lines represent interactions where the AMP atom / group is a donor, and solid lines represent interactions where AMP is an acceptor.**

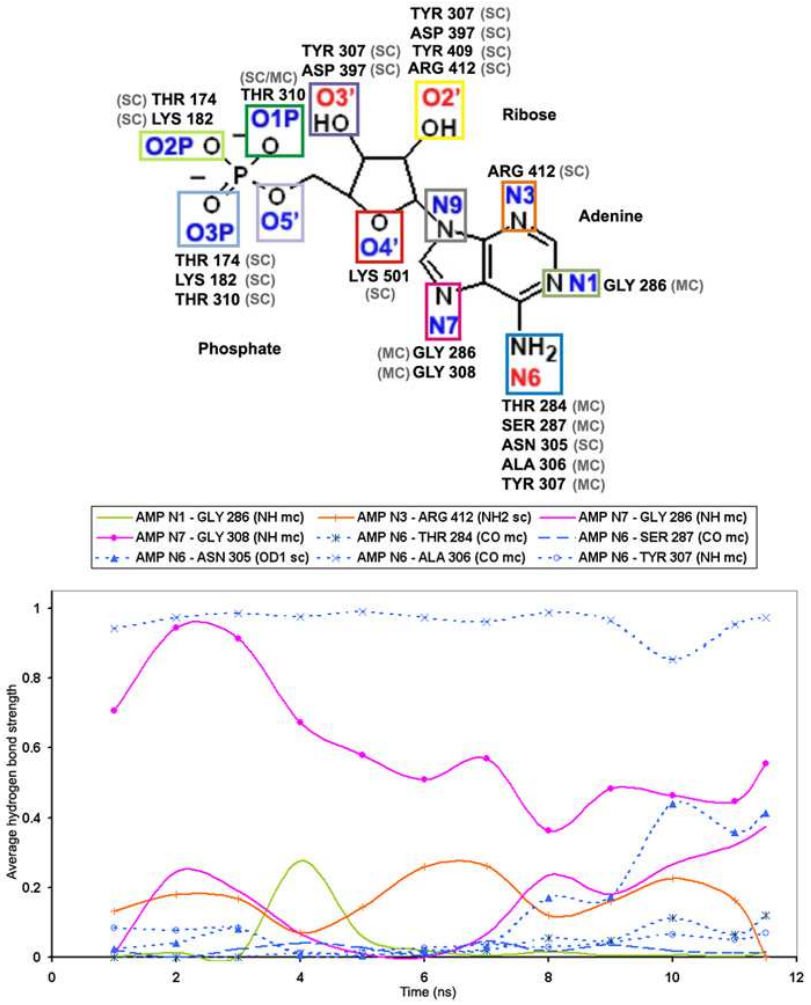


Figure 4.18: **Hydrogen bonding between the adenine moiety of the AMP ligand and PheA in the PheA-Arg simulation. Dashed lines represent interactions where the AMP atom / group is a donor, and solid lines represent interactions where AMP is an acceptor.**

306 main chain carbonyl group is of equivalent strength and duration to that observed in the PheA-Tyr simulation. The interaction observed in the PheA holo and PheA-Tyr simulation between the AMP N7 atom and main chain amino group of Gly 308, was observed in the PheA-Arg simulation however the strength of this hydrogen bonding interaction gradually decreases on the time scale of the simulation.

Ribose Binding

The hydrogen bonding interactions formed between PheA and the ribose moiety of AMP in the PheA-Tyr simulation (shown in the upper graph of figure 4.19), are very similar in strength and duration to those formed in the PheA holo simulations. The hydrogen bonding

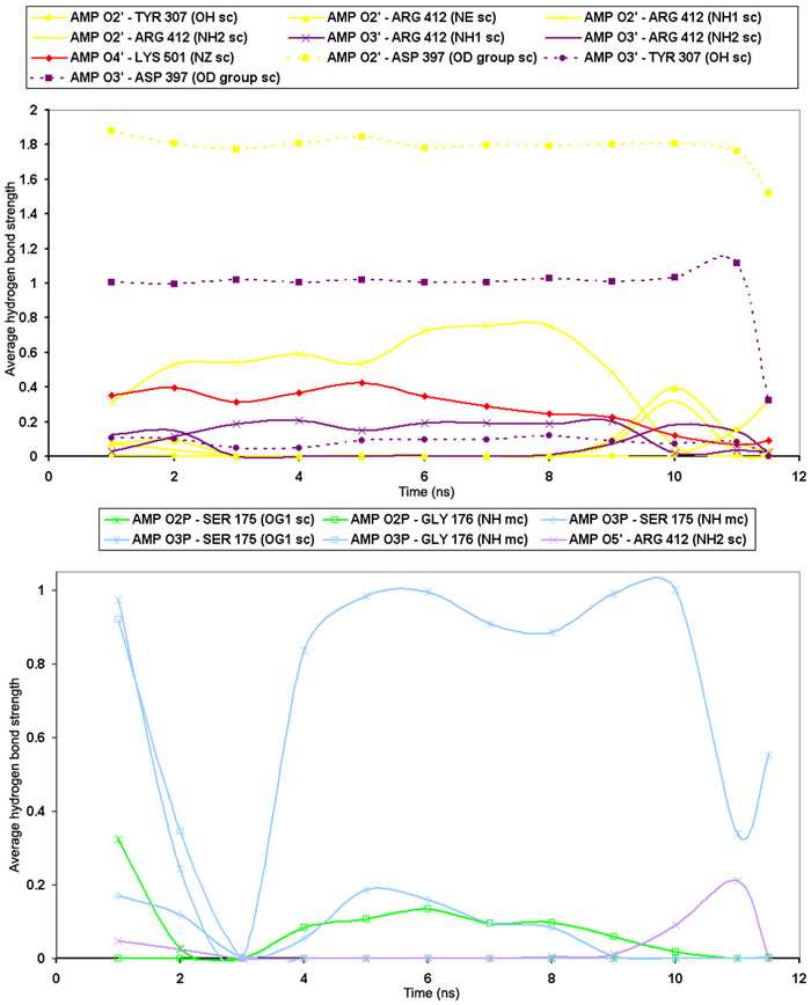


Figure 4.19: **Hydrogen bonding between the ribose (upper graph) and phosphate (lower graph) moieties of the AMP ligand and PheA in the PheA-Tyr simulation. Dashed lines represent interactions where the AMP atom / group is a donor, and solid lines represent interactions where AMP is an acceptor.**

interaction between the 2' hydroxyl sugar of ribose and the Asp 397 sidechain group is strong, 1.8, on the time scale of the simulation. The hydrogen bonding between the 3' hydroxyl sugar of ribose and the Asp 397 sidechain is also strong, 1.0, until the final 500 ps of the simulation when the strength of this hydrogen bonding interaction is reduced to 0.3. The hydrogen bonding of the O4 atom of AMP to the amino group of Lys 501 is slightly weaker (circa 0.4) than that observed in the PheA holo simulations (circa 0.6).

The hydrogen bonding interactions between PheA and the ribose moiety of AMP in the PheA-Asp and PheA-Arg simulations (see the upper graphs of figures 4.19 and 4.21, respectively) are very similar. In both simulations, the hydrogen bonding between the 3' hydroxyl of ribose and the sidechain group of Asp 397 is stronger than that observed in the

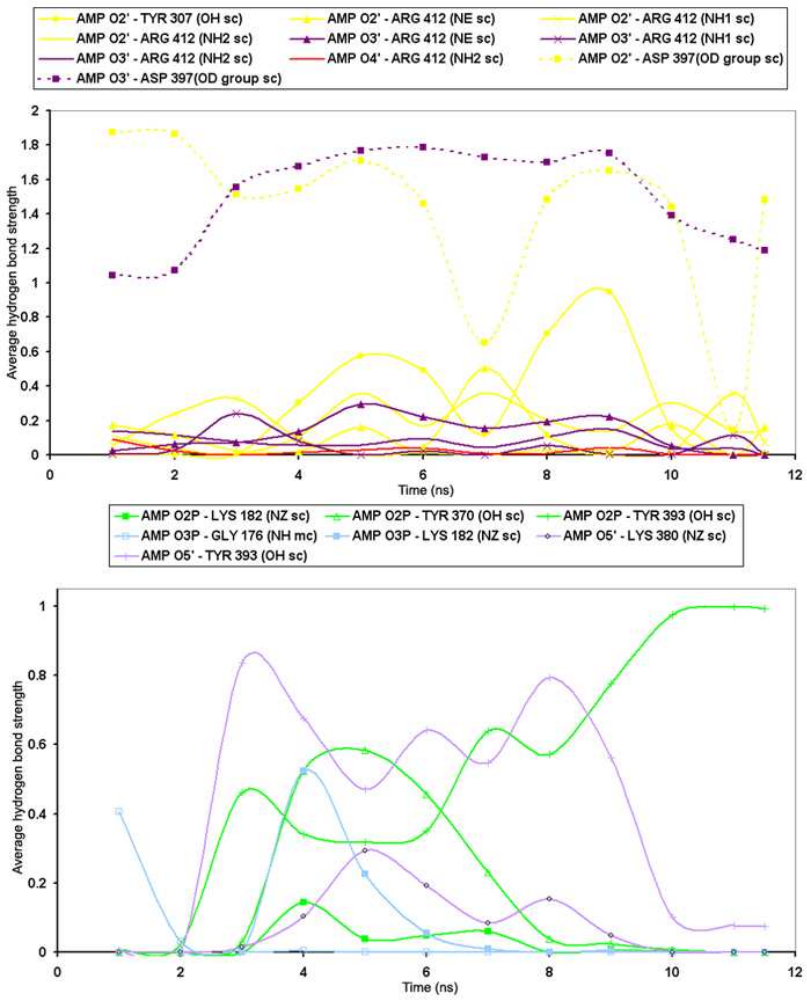


Figure 4.20: **Hydrogen bonding between the ribose (upper graph) and phosphate (lower graph) moieties of the AMP ligand and PheA in the PheA-Asp simulation. Dashed lines represent interactions where the AMP atom / group is a donor, and solid lines represent interactions where AMP is an acceptor.**

PheA holo and Tyr simulations, 1.0–1.8 as compared with 0.8–1.2. Conversely the strength of the hydrogen bonding between the 2' hydroxyl of ribose and the sidechain group of Asp 397 is weaker and varies more on the time scale of the simulation for PheA-Asp and PheA-Arg, than for the PheA holo and Tyr simulations.

Phosphate Binding

The lower graph of figure 4.19 describes the interactions formed between the phosphate group of AMP and the PheA protein in the PheA-Tyr simulation. The primary interaction of PheA with the phosphate group of AMP in the PheA-Tyr simulation is between the O3P oxygen atom and the sidechain hydroxyl group of Ser 175. A hydrogen bond is present

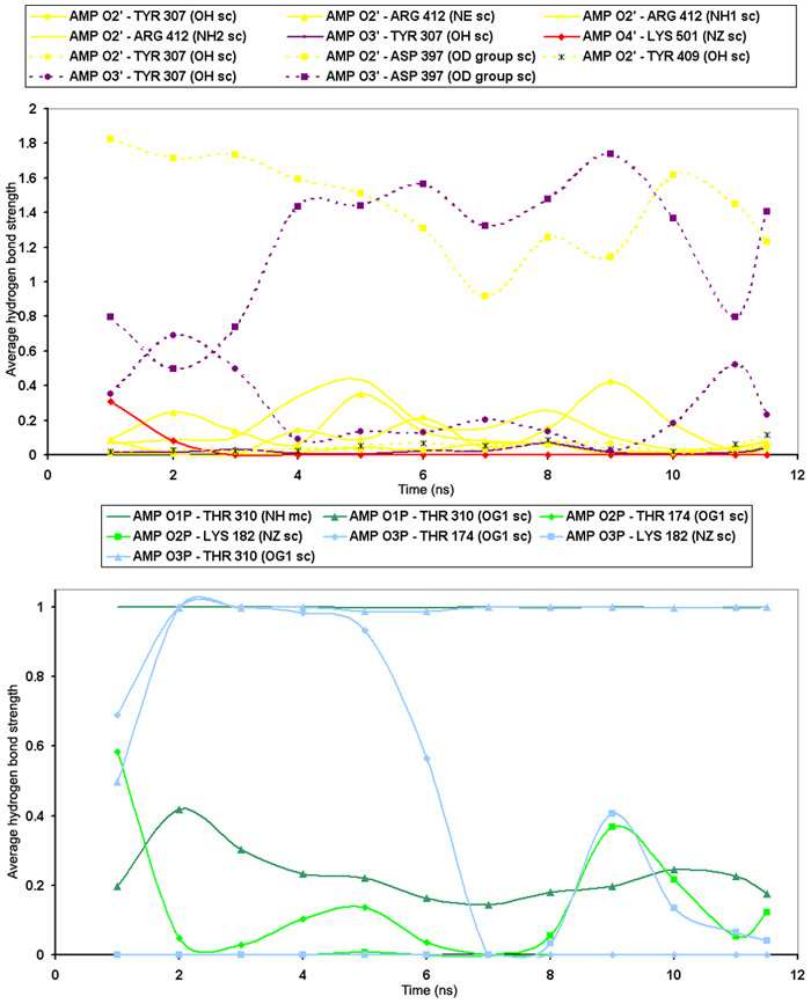


Figure 4.21: **Hydrogen bonding between the ribose (upper graph) and phosphate (lower graph) moieties of the AMP ligand and PheA in the PheA-Arg simulation. Dashed lines represent interactions where the AMP atom / group is a donor, and solid lines represent interactions where AMP is an acceptor.**

during the first nanosecond of the simulation, but the strength of the interaction decrease during the second and third nanoseconds. Between the fourth and tenth nanoseconds the strength of this hydrogen bonding interaction is between 0.8 and 1, and the interaction weakens in the final 1.5 ns of the simulation.

The residues interacting with the phosphate moiety in the PheA-Asp simulation (lower graph of figure 4.20) are quite different to those observed in the PheA-Tyr simulation. Between the third and ninth nanosecond a relatively strong hydrogen bonding interaction (0.5–0.8) is formed between the O5 AMP atom and the hydroxyl sidechain of Tyr 393. From the sixth nanosecond onwards the strength of the hydrogen bonding between the O2P atom of AMP an Tyr 393 hydroxyl sidechain increases from 0.4 to 0.95.

The hydrogen bonding interactions between the phosphate moiety and PheA in PheA-Arg, see figure 4.21, most closely resemble those seen in the PheA2-holo simulation, although not as many interactions are observed in the PheA-Arg simulation. In PheA-Arg, strong hydrogen bonding interactions are formed between: the Thr 310 main chain amino group and O1P atom of AMP - strength 1 for the duration of the simulation, the Thr 310 sidechain hydroxyl group and O3P AMP atom - strength 1 for the final 8.5 nanoseconds of the simulation, and the sidechain hydroxyl group of Thr 174 with O3P AMP - hydrogen bonding of strength 1 during the third, fourth and fifth nanoseconds.

4.4.11 Mg Coordination

The coordination of the magnesium ion in the PheA-Tyr, PheA-Asp, and PheA-Arg simulations is described in appendix figures 7.13, 7.14 and 7.15 respectively.

In each system the Magnesium ion is coordinated to six oxygen atoms. Four oxygen ligands are common to each of the magnesium coordination complexes in the noncognate ligand simulations; the OE1 and OE2 atoms of Glu 311 (pdb: 327), and O1P and O2P oxygen atoms of the AMP phosphate group.

In the PheA-Arg simulation the two further oxygen ligands are provided by water molecules,

as was seen in the PheA2-holo simulation

In both the PheA-Tyr and PheA-Asp the additional two oxygen ligands are provided by the sidechain hydroxyl group oxygens of Thr 174 from the A3 motif and Thr 310.

4.5 Conclusions

The noncognate ligand simulation that displays the greatest similarity in interdomain motion and ligand binding with the PheA holo simulations is the PheA-Tyr simulation. The relative motion of the A_{sub} domain to the A_{core} domain identified by DynDom from the extreme projections of the first two eigenvectors of PheA-Tyr is most similar to that observed in the PheA1-holo simulation; however in the PheA-Tyr simulation domain motion occurs between the full A_{sub} and A_{core} domain, whereas in the PheA1-holo simulation the motion described by the first eigenvector occurs between the subdomain E and helix H6 of the A_{sub} domain and the A_{core} domain and subdomain D of the A_{sub} domain. In both the PheA-Tyr and PheA1-holo simulations flexibility of the A3 motif loop is observed.

Far fewer hydrogen bonds are formed at the left and right sides of PheA and between residues of the A10 motif K loop and the A_{sub} domain in PheA-Tyr than in either the Phe1-holo or PheA2-holo simulation. The hydrogen bonding interactions formed between the L-Tyr substrate and PheA most closely resemble those from the PheA1-holo simulation however interaction of the Lys 501 amino group with the α-carboxyl group of L-Tyr is weaker than the equivalent interaction observed in the PheA1-holo simulation. This is likely due to the size difference between the L-Phe and L-Tyr substrates, with the L-Tyr substrate not being as easily accomodated in the L-Phe binding pocket of PheA .

While the PheA A domain is known not to select the L-Tyr substrate, this MD simulation suggests that interaction between the substrate and Asp 219 (pdb: 235) and Lys 501 (pdb: 517) are a necessary for the interdomain rotation, which as suggested by the results from the cognate holo simulations in Chapter 3, may increase the access pathway between the enzyme domains through which the PPant arm can access the enzyme active site. The

differences observed in domain between which motion occurs in the PheA-Tyr and PheA cognate holo simulations may be due to the larger L-Tyr substrate which does not form as strong hydrogen bonding interactions with the PheA Lys 501 (pdb: 517) residue.

No concerted domain motion between the A_{sub} and A_{core} domains was identified from the PheA-Asp and PheA-Arg simulations. The pattern of interdomain hydrogen bonding in these simulations however is quite similar to that observed in the PheA apo simulations where interdomain motion is observed. Hydrogen bonding between the L-Asp substrate and the Asp 219 residue of the PheA protein is initially present in the PheA-Asp simulation however these interactions weaken during after the first three nanoseconds of the simulation. The L-Asp substrate leaves the PheA binding pocket which most likely occurs as a result of the lack of suitably positioned residues for the L-Asp sidechain to form hydrogen bonding interactions with. This observation is consistent with the observations of Ackerley and co-workers and Lautru and co-workers who suggest, from analysis of the substrate specificity code, that smaller substrates are thought to utilise only the residues at the top of the biding pocket [30,85]

On the timscale of the simulation, the L-Asp substrate α-carboxyl group forms a number of hydrogen bonds with the A3 motif loop residues as it exits the binding pocket. These observations suggest the A3 motif loop may have a role in facilitating the removal of noncognate ligands from the binding site.

4.5.1 Summary of Domain Motion

Figure 4.22 represents the motion observed between the A_{core} and A_{sub} domain in the PheA-Tyr and PheA-Asp domains.

In PheA-Tyr the largest motion occurs at 10ns with the A_{sub} domain moving towards the A3 motif loop and exposing and widening the Ppant active site. Prior to this, at 8.6ns, the A_{sub} domain tips towards the A3 motif loop reducing access to the Ppant active site. The A3 motif loop is flexible on the timescale of the simulation.

In the PheA-Asp simulation concerted domain motion is not observed, however on the timescale of the simulation the L-Asp substrate does leave the active site. This is accompanied by a tilting of the A_{sub} domain away from the interdomain interface, opening the protein and lifting the A10 motif loop (shown in blue in figure 4.22), and an opening of the A3 motif loop, shown in red in figure 4.22. This lifting of the A10 motif loop and A_{sub} domain is in line with the open A domain conformation observed in the SrfA-C A domain.

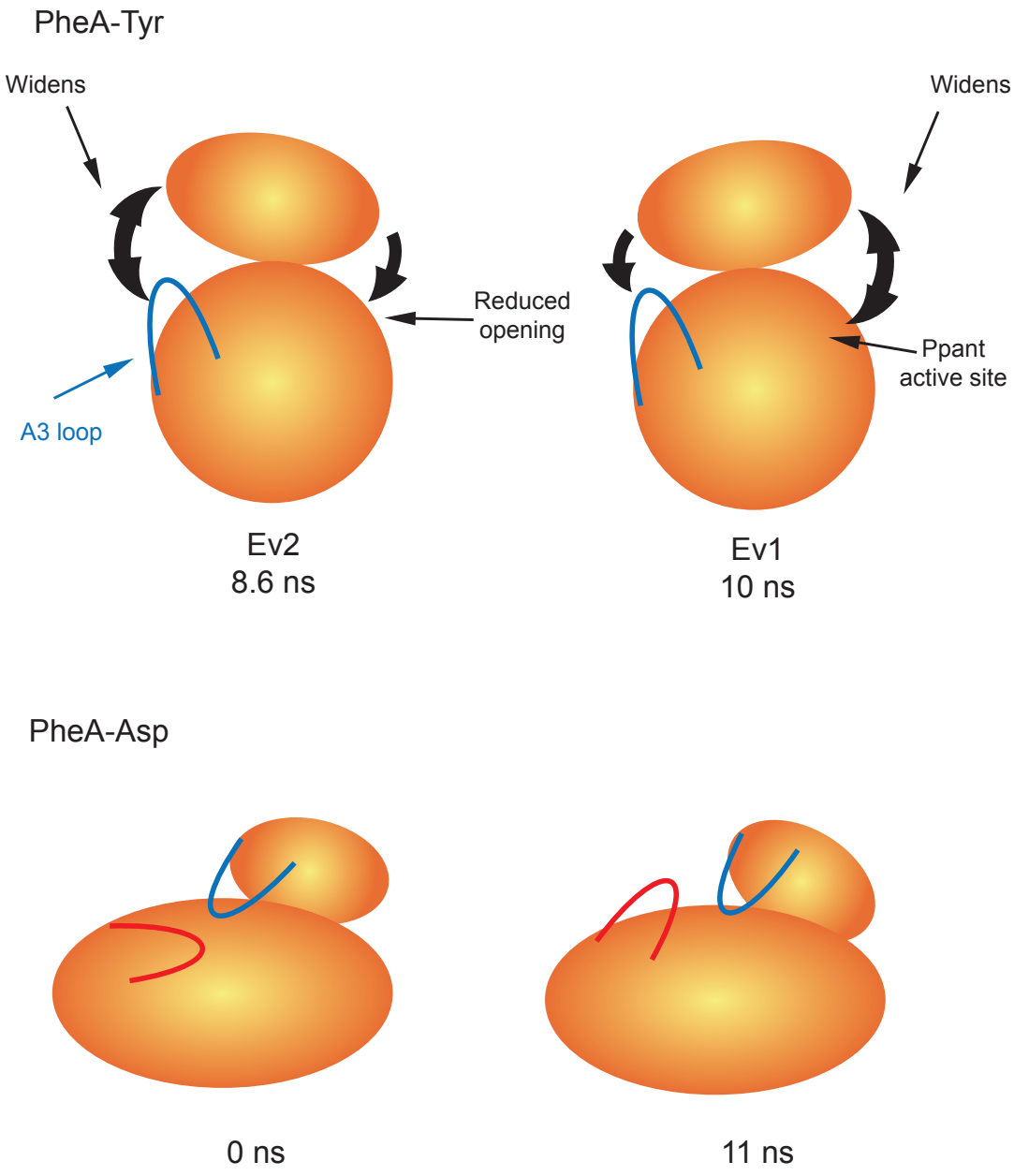


Figure 4.22: **Schematic of the domain motion observed in the PheA-Tyr and PheA-Asp simulations.**

Chapter 5

Molecular Modelling of the Module 2 Adenylation Domain, CchH2, from *Streptomyces Coelicolor*

5.1 Overview

In this chapter the results of a homology model of the second A domain of the NRPS that forms Coelichelin are presented[23]. This is an iterative NRPS and so it is of interest to understand whether the behaviour of the A domains in these types of NRPSs differ from non-iterative NRPSs. As there is no structure for the A domains from this NRPS, a homology model was built for the A domain using PheA as a template. MD simulations were carried out with the apo structure, cognate and noncognate substrates. Analysis of the RMSD and secondary structure of the homology model suggests the model is stable on the timescale of the simulations, despite the low sequence identity with the template structure. Slightly less stability is observed for some α-helices as compared to the PheA cognate substrate holo simulations however these α-helices are not considered part of the core A domain structures. Differences are also observed when comparing the RMSD of the α-helices and β-sheets of the A_{core} domain of CchH2 with the equivalent region from the PheA-holo simulations presented in Chapter 3. No interdomain motion is observed in the CchH2-apo simulation or the CchH2-Val simulation. The first eigenvector from the CchH2-Thr and CchH2-Ser simulation, and second eigenvector from the CchH2-Ser simulation describe interdomain motion. The CchH2 holo simulations show that the cognate substrate L-Thr forms stronger hydrogen bonding interactions with the A domain than the noncognate substrates. These results suggest that homology modelling of the A domains may be a useful technique for further study of the dynamics and substrate interactions of the A domains.

5.2 Introduction

Members of the actinomycetes, particularly those from the *Streptomyces* genus, produce in excess of 65% of the known microbial antibiotics and a number of other commercially important pharmaceuticals and agrochemicals[281]. Compounds produced by the *Streptomyces* genus encompass the majority of natural product classes, including β-lactams, oligosaccharides, terpenes, peptides, polyketides and alkaloids. Genome sequencing of the most

thoroughly characterised member of this genus, *Streptomyces coelicolor*, began in 1999[282] and was completed in 2002[283].

5.2.1 Coelichelin

In 2000, an NRPS homologue was identified on cosmid SCF-34 of the *S. coelicolor* ordered genomic library. This gene, named *cchH*, encodes a protein, CchH, consisting of 3643 amino acids with a predicted molecular mass of 390 kDa. Sequence analysis of CchH applying the conserved sequence motifs for NRPS domains[22], identified ten domains; three Adenylation (A) domains, three peptidyl carrier protein (PCP) domains, two Condensation (C) domains, and two epimerisation (E) domains, arranged within three modules (see figure 5.1a). Unusually no thioesterase (Te) or reductase (Red) domain is present at the C-terminal end of module 3. The CchH synthetase, therefore, contains no domains capable of releasing the assembled peptide product into the solution[284].

The substrate specificity of the CchH A domains was predicted using the Challis-Ravel specificity conferring code[285]. CchH modules 1, 2 and 3 were predicted to recognise and selectively activate L-δ-*N*-formyl-δ-*N*-hydroxyornithine (L-hfOrn), L-threonine and L-δ-*N*-hydroxyornithine (L-hOrn), respectively[284] (see figure 5.2). The well-documented preference of NRPS A domains for L-amino acid substrates[30] and application of the 'colinearity rule', coupled with the presence of epimerisation domains in modules 1 and 2 of CchH, suggested a product with a D-D-L configuration and thus two structures for the putative tripeptide product were proposed[284], structures **c1** and **c2** in figure 5.1. Of these two structures, **c2** seemed to be the likely structure, as cleavage of **c1** from the synthetase would require a terminal Te domain, unnecessary for the release of structure **c2**. Instead cleavage of structure **c2** from CchH2 was proposed to occur via attack of the 5-amino group of the C-terminal 5-hydroxornithine residue through a kinetically favoured 6-*exo*-trig transition state[286].

The preferred proposed structure, **c2**, for the CchH product is similar to that of the siderophores produced by *Mycobacterium smegmatis* and *Mycobacterium neoaurum*, exochelin MS and

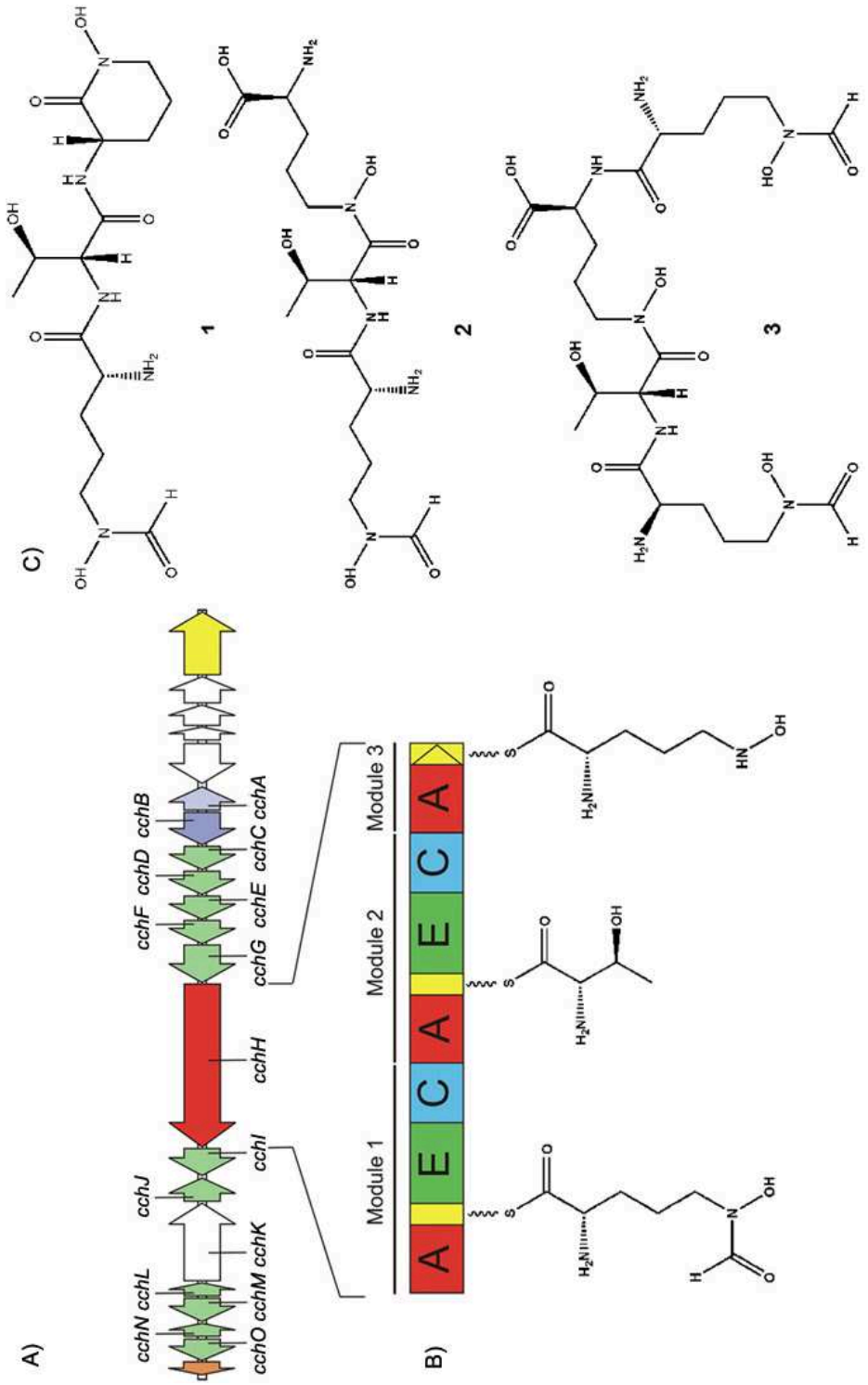


Figure 5.1: **Organisation of the coelichelin, *cch*, biosynthetic gene cluster and NRPS. A:** The *cch* gene cluster. Genes are coloured as followed: chitinase (*chiG*): orange; regulation, transport and degradation genes: green; genes of unknown function: white; peptide synthetase: red; L-ornithine 5-monooxygenase (*cchB*): mauve; formyl transferase (*cchA*) : lilac; RNA helicase: yellow. **B:** The ten domains that comprise the CchH NRPS; three Adenylation (A) domains, three peptidyl carrier protein (PCP) domains, two Condensation (C) domains and two epimerisation (E) domains; arranged into three modules. The relevant substrates are attached to each modules PCP domains. **C:** 1 and 2 are the tripeptide structures suggested to be the hypothetical product of CchH. Structure 3 is the experimentally determined tetrapeptide Coelichelin. Illustration adapted from figure 1[23].

A domain[a]	Residue (according to GrsA numbering) 235	236	239	278	299	301	322	330
FxbB-M1-hfOrn	D	I	N	Y	W	G	G	I
CchH-M1-hfOrn	D	I	N	Y	W	G	G	I
Cda1-M2-Thr	D	F	W	N	V	G	M	V
AcmB-M1-Thr	D	F	W	N	V	G	M	V
SnbC-M1-Thr	D	F	W	N	V	G	M	V
FenD-M2-Thr	D	F	W	N	I	G	M	V
SyrB-M1-Thr	D	F	W	N	L	G	M	V
SyrE-M7-Thr	D	F	W	N	V	G	M	V
PvdD-M1-Thr	D	F	W	N	I	G	M	V
PvdD-M2-Thr	D	F	W	N	I	G	M	V
FxbC-M2-Thr	D	F	W	N	I	G	M	V
CchH-M2-Thr	D	F	W	N	I	G	M	V
FxbC-M1-hOrn	D	M	E	N	L	G	L	I
FxbC-M3-hOrn	D	M	E	N	L	G	L	I
CchH-M3-hOrn	D	M	E	N	L	G	L	I

Figure 5.2: **Prediction of the substrate specificity determining residues and amino acid substrates of the CchH A domains.** [a] Nomenclature: protein name-module number-amino acid substrate (hfOrn: L-δ-*N*-formyl-δ-*N*hydroxyornithine, Thr: L-threonine, hOrn: L-δ-*N*-hydroxyornithine). Table adapted from table 1 [284].

exochelin MN respectively. This, in addition to the suggestion that the threonine hydroxyl group and the two hydroxamic acid groups are potential Fe^{3+} chelators, lead to the proposal that the CchH product is a novel *S. coelicolor* siderophore. The product was named Coelichelin [284]. Analysis of the genes surrounding *cchH* showed it was part of a cluster of 15 genes (*cchA*−*cchO*) that span 29 kb of cosmid SCF-34. The genes in the *cch* cluster (see figure 5.1a) were suggested to be involved in the synthesis, transport, and degradation of coelichelin. The function of these genes was deduced based on their homology to genes of known function [284].

Efforts to isolate, determine the composition and sequence of amino acids in and the total structure of coelichelin were achieved using a combination of experimental and computational techniques. The experiments revealed coelichelin to be a tetrapeptide assembled by a trimodular NRPS. The production of coelichelin by CchH is therefore an example of non-linear nonribosomal peptide synthesis. The functioning of CchH as a non-linear NRPS was something that could not have been predicted from sequence analysis as it was without literature precedent [23].

The identification of the product of CchH as a siderophore and the proposal of two potential structures was critical for identifying the experimental conditions under which the *cch*

cluster is expressed; expression of a siderophore producing gene cluster would occur in an iron-deficient environment. A combination of gene knockouts and metabolic profiling was used to identify coelichelin. A mutant strain of *S. coelicolor* M145, *S. coelicolor W5*, was produced by inactivating *cchH*. Comparative HPLC analysis of culture supernatant of these two *S. coelicolor* strains grown under iron-deficient conditions, identified a compound which forms a complex with ferric iron that is present in strain M145 but lacking in strain W5. Production of this compound by *S. coelicolor* M145 was suppressed by the addition of ferric iron to the culture medium. The maximum absorption (λ_{max}) of 435 nm suggested ferri-coelichelin was a *tris*-hydroxamate complex. Semipreparative HPLC was used to partially purify ferri-coelichelin from the *S. coelicolor* supernatant. Desferri-coelichelin was purified to homogeneity by a further semi-preparative HPLC step after removal of the ferric iron by treatment with 8-hydroxyquinoline[23].

The identity of the amino acid residues and their sequence within coelichelin was determined using a combination of high resolution and tandem mass spectrometic analysis of desferri-coelichelin and one- and two- dimensional high-field NMR analysis of gallium-coelichelin. Molecular modelling guided by experimental data, specifically inter-residue distances and dihedral angles calculated for Ga-coelichelin from the ROESY and ^{1}H NMR data respectively, was used to determine the relative stereochemistry of the four coelichelin α-carbon atoms. Acid-promoted hydrolysis of Ga-coelichelin followed by conversion of the liberated threonine to its N-trifluroacetyl isopropyl ester derivative and comparison by chiral gas chromatography with authentic standards, was used to determine the relative configuration of the α- and β-carbon atoms of threonine[23].

These results, the arrangement and type of domains present within the CchH NRPS and the preference of NRPS A domains for L-amino acid substrates lead to the conclusion that the absolute stereochemistry of coelichelin is D-hfOrn-D-*allo*-Thr-L-hOrn-D-hfOrn. The total structure of the *tris*-hydroxamate tetrapeptide coelichelin, **c3**, is shown in figure 5.1[23]. The structure of coelichelin is very similar to the favoured predicted structure, **c2**, although the **c2** structure lacks the second D-hfOrn residue[23].

The authors' use of the Challis-Ravel model for substrate specificity prediction[285], as op-

posed to the alternate Stachelhaus method[81], ensured accurate prediction of the substrates combined into the CchH product, coelichelin, as this method discriminates between A domains that activate Orn, hOrn and hfOrn[285 284].

CchH was shown to be the only NRPS responsible for coelichelin biosynthesis by expressing the *cch* cluster in the heterologous host *Streptomyces fungicidicus* B-5477. Testing for coelichelin production in this mutant and wild type *S. fungicidicus* cultures grown in iron-deficient medium was done by comparative HPLC analysis. The mutant strain was shown to produce significant quantities of coelichelin whereas wild-type *S. fungicidicus* did not produce coelichelin[23].

The release of coelichelin from CchH was determined to be facilitated by the enterobactin esterase homolog CchJ. Replacing *cchJ* on the chromosome of *S. coelicolor* M145 to generate *S. coelicolor* W6, resulted in suppression of coelichelin production when this mutant was grown under iron deficient conditions. These results are consistent with CchJ acting as a thioesterase that hydrolytically releases the tetrapeptidyl thioester from the module 3 CchH PCP domain[23].

The synthesis of a tetrapeptide by a trimodular NRPS raises intriguing questions about the mechanism of the peptide product assembly. The proposed mechanism for coelichelin biosynthesis suggests that after one complete elongation cycle, module 1 of CchH (CchH1) is 're-used', incorporating a second molecule of the substrate L-hfOrn into the final peptide product. The domains in CchH are therefore used iteratively; in the first elongation cycle each of the first three substrates is incorporated and in the second cycle a second molecule of substrate L-hfOrn is incorporated by module 1, and a number of domains from the second and third modules are skipped[23].

The identification of the *cchH* gene cluster that encodes the CchH NRPS, prediction of the A domain substrate specificity using the Challis-Ravel specificity conferring code[285] and experimental determination of the product structure, illustrated and validated an efficient novel method for predicting the structure of unknown natural products directly from genome sequence data.

5.2.2 Molecular Modelling of the CchH A domains

To date the PDB[287] contains the structures of A domains in complex with their respective substrates, and in the PheA and DhbE structures Adenosine Monophosphate (AMP) is bound. The structure of DhbE has additionally been determined in the adenylate state, in a complex with DHB-AMP, and in the apo state.

As there are very few structural data for the A domains any further studies of the dynamics and the molecular mechanism of the substrate specificity of the A domains, as were carried out with PheA in Chapters 3 and 4, require the building of structural models using theoretical methods. The method that generally produces the most accurate models is comparative modelling[189]. The majority of protein sequences are detectably related to known protein structures by less than 30% sequence identity. For this reason comparative modelling is routinely performed on target and template sequences which share less than 30% sequence identity[183]. Models with a quality comparable to that of those determined experimentally can be produced providing the identity of the two sequences is sufficiently high[185]. Significant errors are commonly observed in alignments between a pair of detectably related sequences that share less than 30% sequence identity[183 288]. At 30% sequence identity approximately 20% of residues in an alignment between two related sequences will be misaligned[289].

The accuracy of predictions made using homology modelling are largely based on the accuracy of the sequence alignment on which the model is built, the quality of which in turn depends upon the level of sequence identity between the template (known structure) and target sequences (model structure)[185]. The CchH A domains have a low sequence identity with the structure of the α-amino acid activating A domain PheA. Additionally, the substrates of CchH1 and CchH3 are not common amino acids and there are no pre-existing force field parameters for these molecules in the GROMOS96 force field ffG43a2[233] which has been used for the MD simulations presented in this thesis. For these reasons, CchH2 - the module two A domain which specifically selects and activates L-threonine - was chosen as the protein to initially study using molecular modelling techniques.

The aims of this study were; to build a CchH2 homology model, carry out docking simulations of CchH2 with the native threonine substrate and the similar yet non-native ligands serine and valine, and to perform MD simulations of the solvated apo structure and three holo systems (CchH2, substrate, AMP and $^{2+}$). The findings of these simulations could be used to provide insight into how CchH2 discriminates between chemically or volume similar substrates and observe the effect of the substrates on the dynamics of the A domain. Depending on the findings of this initial study, it was hoped the methodology implemented within could eventually be expanded to model the CchH1 and CchH3 A domains and the interactions with their respective substrates, L-hfOrn and L- hOrn.

5.3 Methods

5.3.1 Homology Modelling

The protein sequence for the NRPS system CchH located on SCF 34.11c (SCO0492) of the *S. coelicolor* ordered cosmid library was obtained from TrEMBL (Q9RK14). The three CchH A domains were located and extracted using an A domain hidden Markov model (HMM) which was developed by the author of this thesis.

The crystal structure of the gramicidin synthetase A, GrsA, adenylation domain, PheA, (pdb: 1AMU) was used as a template for the construction of the CchH2 homology model. The sequences of GrsA and CchH2 share $\sim$30% identity. DhbE was not selected as the primary template on which to model CchH2 because these two sequences share less than 30% sequence identity and because DhbE is a freestanding A domain activating a non α-amino acid. As the PheA crystal structure is lacking residues from the AMP binding loop (Pfam[290]: PF00501) the PheA structure used for the MD studies in Chapters 3 and 4 was used as a template for the modelling of CchH2.

Using the align2d class of MODELLER v 8.1[190], an initial alignment of PheA (the template) and CchH2 (the target) was generated. This alignment was then evaluated using the MODELLER alignment.check routine which checks for two criteria. The first applies only

to cases where more than one template structure has been used; each pair of aligned C-α atoms is superimposed and any atoms that are more then 6 Å away from each other are identified. Pairs of C-α atoms more then 6 Å apart are almost certainly misaligned. The second criteria checks the alignment of the target sequence with each of the templates; the distance between each consecutive pair of C-α atoms in the template is checked by measuring the distance between the equivalent consecutive pair of C-α atoms in the target. Consecutive C-α atoms in the template which correspond to atoms in the target greater than 8 Å apart are flagged as these regions of alignment (predominantly associated with gap insertions) are almost certainly incorrect.

Using the MODELLER alignment 100 models of CchH2 were built based on the satisfaction of spatial restraints. As the structure of PheA was determined in the presence of AMP which is common to both PheA and CchH2 and as CchH2 is predicted to bind this ligand in a similar manner, based on sequence similarity and motifs, all models of CchH2 were built including these ligands. The best model was chosen by comparing the MODELLER objective functions (MOFs) from each model; the lower the MOF the better the model. Energy profiles for this model were generated using Prosa2003[200] and the MODELLER Discrete Optimized Protein Energy (DOPE) routine. These energy profiles were compared with that of PheA to identify problematic regions within the theoretical model. Additional assessments of the model were carried out using the MODELLER ga341 (Z-score and energy analysis) and superpose routines (RMSD between PheA and CchH2 model), Prosa2003 to perform a Z-score, and energy analysis, and Procheck which assesses the stereochemical properties of the model structure. Parameter and run files for Modeller are included in appendix IIII.

Comparison of the PheA and CchH2 Prosa2003 and DOPE energy profiles highlighted some regions of high energy structure in the CchH2 model which were not observed in the equivalent regions of PheA. As the majority of these high energy regions corresponded to the location of insertions in the alignment, alternate options for inserting these gaps into the alignment were identified by aligning the two sequences using the alignment programs; Clustal, T-Coffee[291], and Muscle[292] and the fold prediction programs; 3DPSSM[293,294],

Phyre[294,295], and mGenTHREADER[296,297].

The secondary structure of CchH2 was predicted using the Phyre server which combines the predictions of PSIPRED[298], JNet[299] and SSpro[300] to produce a consensus prediction. Regions of appropriate secondary structure were only assigned to CchH2 if they were predicted with a confidence above 6. The seven alignments (MODELLER, Clustal, T-Coffee, Muscle , 3D-PSSM, Phyre, and mGenTHREADER) were annotated with the secondary structural prediction for CchH2 and the location of secondary structural elements in PheA, as defined by Conti *et al.*[62], which was retrieved from the PDB.

The viability of the location of the insertions in the alignments was judged based on the criteria in the list below and either accepted, rejected or refined by manual alignment.

1. Correct alignment of secondary structure elements
2. Insertion falling outside regions of secondary structure
3. Correct alignment of residues in substrate binding pockets
4. Correct alignment of residues in the AMP binding pocket
5. Correct alignment of residues in A domain motifs[22]
6. Correct alignment of residues in the AMP binding motif (PF00501)
7. The insertion creates no distance violations (MODELLER check routine)

The effect of altering the position of the gap insertions in the original MODELLER alignment on the quality of the model produced was tested in an iterative fashion. For each new alignment produced 100 homology models were built, the best model determined, and analysed. Regions in the model determined by Prosa2003 to be of high energy and corresponding to a region of gap insertion in the alignment were then targeted for alteration in the subsequent iteration of this process. This iterative procedure, that utilised the informed use of the secondary structure annotated alignments, continued until no further improvements were seen in the model quality by altering the alignment.

All of the CchH2 models shared a common region of high energy, identified by DOPE and Prosa2003, not shared by the PheA structure. Visual assessment of the alignment showed no insertions affecting this region but from the secondary structure annotation it was clear that CchH2 was predicted to contain a significantly longer helix than that found in PheA. The length of the CchH2 helix was comparable to that found in the equivalent region of DhbE, the freestanding A domain from *Bacillus subtilis*. Based on these observations, a new alignment was manually generated using the appropriate region of DhbE as a template on which to model the longer CchH2 helix. Using this alignment and both PheA and DhbE as templates, 100 models were built. The best model was selected and analysed, as previously described.

Loop regions within this model identified as regions of high energy structure by Prosa2003 and therefore with the potential for refinement, were sequentially optimised using the loop-model class in MODELLER[267] to produce 100 alternate loop conformations for each selected region. After each loop region was optimised and the best loop selected, based on the MOF, the effect of this new loop on the overall model quality was assessed, as previously described and the new model accepted or discarded.

5.3.2 Docking

Hydrogen atoms were added and energy minimisation performed on the CchH2 complex prior to docking as is described in the section 5.3.3. The AutoDock 3.0.5 program[220] with the Lamarckian genetic algorithm was used for automated docking. AutoDockTools (ADT) was used to process files for input to the autogrid3 and autodock3 routines of AutoDock. Prior to performing the docking runs the individual constituents of the macromolecule (CchH2, AMP and Mg^{2+}) were prepared. ADT was used to merge the non-polar hydrogen atoms, assign Kollman charges and solvation parameters to the CchH2 structure. Gasteiger charges and solvation parameters for AMP were those used in the simulations in Chapter 4. A charge of +2 and solvation parameter of 0 was assigned to the Mg^{2+}.

Structures of the substrates (L-Thr, L-Ser and L-Val) were obtained using Quanta. ADT

was used to merge the non-polar hydrogen atoms, add Kollman charges, determine the rigid root of the ligand, and the number of active torsions. The docking parameters used were consistent with those used in Chapter 4. The docking grid was set to contain 90 x 90 x 90 points with a grid spacing of 0.0275 nm. The grid was placed symmetrically about the centre of mass of the molecule. This ensured the grid boxes included the entire enzyme binding site and also provided enough space for the ligand translational and rotational walk. After the grid parameter file was written, it was edited manually to contain the appropriate parameters for the phosphorous and magnesium atoms as are defined in the AutoDock documentation. The grid parameter file was used to generate the docking grid by running autogrid from within ADT.

Next the docking parameters were defined, the docking parameter file created and AutoDock run from within ADT. For each enzyme-substrate (ligand) complex 100 runs were performed. For each run a maximum number of 25,000 genetic algorithm operations were generated on a single population of 50 individuals. The maximum number of energy evaluations was set to 250,000. The ligand was placed in a random starting position and conformation at the beginning of each docking run. During the run the ligand was allowed a maximum mutation of 0.2 $\AA$ in translation and 50 $^\circ$ in rotation. Parameters specific to the Lamarckian genetic algorithm include a mutation rate of 0.02, a crossover rate of 0.8, elitism of 1 and a local search rate of 0.06.

The docking simulation results were ranked according to the docked energy between the protein and the ligand, a summation of internal ligand energy and intermolecular energy terms. A conformational clustering analysis was performed on the resulting docked structures. The orientation of the top ranking structure in each cluster was visualised in relation to CchH2, AMP and Mg^{+2} using VMD. The orientation of the ligand in the binding pocket of CchH2 was compared to that of L-Phe in the binding pocket of PheA. The CchH2 docked ligand which had an orientation most consistent with that observed in the PheA crystal structure (pdb: 1AMU) and made the required contacts with the Asp and Lys binding pocket residues was selected to be used as the starting structure for the MD simulations.

5.3.3 MD Simulations of the CchH2 Homology model

Summary of Simulations

Four simulations were performed: one apo state CchH2 simulation and three holo state CchH2 simulations each containing the cofactors (AMP and Mg^{2+}) and one of each of the following substrates; L-Threonine, L-Serine and L-Valine. The apo CchH2 simulation was performed in a truncated octahedral box, 747.13 nm^3 which when solvated was filled with 23026 water molecules. Each holo CchH2 simulation was performed in a truncated octahedral box, 770 nm^3. The simulation with L-Thr (CchH2-Thr) and L-Ser (CchH2-Ser) substrates both contained 23020 water molecules and the simulation with L-Val (CchH2-Val) 23021 water molecules.

Simulation parameters

All simulations were performed using the GROMACS 3.2.1 simulation suite of programs (www.gromacs.org)[235] and the GROMOS96 43a2 united atom force field[233]. Unless otherwise specified, the following parameters were used to perform the MD simulations presented in this chapter.

Polar hydrogen atoms were added to the protein and substrate using the GROMACS pdb2gmx routine. Lysine and arginine sidechains were modelled as protonated residues, aspartic acid and glutamic acid as unprotonated residues and histidine residues as neutral residues. The hydrogen atom was added to each histidine residue in an automated fashion, by the GROMACS pdb2gmx routine, based on the optimal hydrogen bonding conformation. The AMP topology from previous simulations was used (see section 3.2.7 of Chapter 3) and the Mg^{2+} ion parameters supplied by the GROMACS ions.itp file.

Energy minimisation was used to relieve steric conflicts generated during the setup. The convergence criteria for energy minimisation, $g = 0 +/- e$, is when the gradient (g) reaches a value within e of 0. Unless otherwise stated the value of e used was 1000 kJ mol^{-1} nm^{-1} and during restrained runs (energy minimisation or molecular dynamics) specified atoms

were tethered to their original position using a harmonic potential with a force constant of 1000 kJ mol^{-1} nm^{-2}. Each system followed the same generalised protocol: minimisation (first restrained and then unrestrained), restrained MD - to ensure relaxation of the solvent - followed by an 11.5 ns unrestrained run of MD.

The SPC water model[269] was used and the systems solvated using the GROMACS genbox routine. This routine fills the box with multiple equilibrated configurations of 216 SPC water molecules and then removes any water molecules within a distance corresponding to the total van der Waals radius of both the water and solute atoms, of the solute molecule. Overall neutrality was achieved by replacing ten randomly selected water molecules with ten Na^+ ions using the GROMACS genion routine. The particle mesh Ewald (PME)[238 301] method was used, for the treatment of long-range electrostatic interactions, with a 1 nm cutoff for the real space calculation and a Fourier spacing of 0.12 nm. The van der Waals interactions were modelled using a 1 nm cutoff. The restrained dynamics simulations were performed in the constant number of particles, volume and temperature (*NVT*) ensemble. The unrestrained simulations were performed in the *NPT* ensemble; constant number of particles, pressure and temperature. The temperature was maintained at 310 K by separately coupling the protein, solvent (water plus Na^+ counterions) and, when present, the AMP, substrate and Mg^{+2} using a Berendsen thermostat[242] with a coupling constant τ_T of 0.5 ps for all of the holo simulations and the restrained apo simulations, and 1.5 ps for the apo simulation production run. The pressure of the system was coupled isotropically using the Berendsen barostat at 1 bar with a coupling constant τ_P = 1.0 ps and compressibility = 4.5 10^{-5} bar^{-1}. A timestep of 2 fs was employed for all simulations. The centre of mass motion of the entire system was removed at every timestep to maintain the effective simulation temperature of 310 K. During the MD simulations all bonds lengths were restrained using the LINCS algorithm[230]. Coordinates and velocities were saved every 1 ps. Initial velocities were generated at 300 K.

The apo CchH2 simulation was performed on the EPSRC Columbus-lx cluster which consists of 22 Dual Opteron nodes and the holo simulations were performed on a Pentium II linux workstation. All simulations were performed between April 2006 and December

2006.

Docking Minimisation Protocol

Energy minimisation of the CchH2/cofactor complex, prior to use in the docking calculations, was performed in the following stages:

1. Less than 50 steps of steepest descents where all non-hydrogen atoms were tethered.
2. Less than 500 steps of steepest descents, where e = 10 and all non-hydrogen atoms, bar the Mg^{2+}, were tethered.

Apo Minimisation and Restrained Simulation Protocol

All energy minimisations of the apo state CchH2 were performed until the maximum energy derivative was less than e = 100 kJ mol^{-1} nm^{-1} and atoms were tethered using a harmonic potential with a force constant of 1000 kJ mol^{-1} nm^{-2} , unless otherwise specified.

The unsolvated apo state CchH2 system energy minimisation was performed in the following stages:

1. 2000 steps of steepest descents all heavy atom were tethered.
2. 3500 steps of steepest descents un-tethered energy minimisation.
3. 10 steps of an un-untethered conjugant gradients energy minimisation.

After the apo state CchH2 was solvated, but prior to the addition of Na^{+} counterions to neutralise the system, energy minimisation was performed in the following stages:

1. 100 steps of steepest descents with all non-hydrogen atoms tethered.
2. 5000 steps of un-tethered steepest descents.
3. 200 steps of un-tethered conjugant gradients.

After ions were added to the solvated apo state CchH2 system energy minimisation was performed in the following stages:

1. 50 steps of steepest descents with all non-hydrogen atoms tethered.
2. 5000 steps of unrestrained steepest descents
3. 100 steps of unrestrained conjugant gradients

A series of *NVT* molecular dynamics simulations were performed were tethering was applied to various atoms and released gradually. This process was to allow gradual relaxation of both the homology model and solvent. In these simulations all specified tethered atoms were subjected to an isotropic force constant of 1000 kJ/mol^{-1} nm^{-1}. These tethered atom simulations were carried out in the following order:

1. 10ps where all non-hydrogen atoms were tethered.
2. 25ps where all non-hydrogen main chain and binding pocket residue atoms were tethered.
3. 50ps where all non-hydrogen main chain atoms were tethered.
4. 50ps where only the non-hydrogen binding pocket atoms were tethered.
5. 100ps of NVT MD with no atoms tethered.

Holo Minimisation and Restrained Simulation Protocol

Unless otherwise specified energy minimisations progressed until the maximum energy derivative was less than 1000 kJ mol^{-1} nm^{-1} and atoms were tethered using a harmonic potential with a force constant of 1000 kJ mol^{-1} nm^{-2}. The unsolvated holo state CchH2 system was subjected to less than 50 steps of steepest descents where all non-hydrogen atoms were tethered. After the water and Na^+ ions were added to achieve overall neutrality, the system was subjected to:

1. 50 steps of steepest descents where all non-hydrogen atoms were tethered.
2. 50 steps of steepest descents where all non-hydrogen atoms were tethered with a force constant of 500 kJ mol^{-1} nm^{-2}.
3. 50 steps of unrestrained steepest descents.
4. 50 steps of unrestrained conjugant gradients.

Before the 11.5 ns untethered MD simulations two 250 ps *NVT* tethered MD simulations were performed. These were to allow relaxation of the solvent. In the first tethered run all non-hydrogen atoms were subjected to an isotropic force constant of 1000 kJ/mol^{-1} nm^{-1} which was reduced to 500 kJ/mol^{-1} nm^{-1} in the second *NVT* run.

MD Analysis

Analyses of the MD trajectories were performed as described in chapter 3 section 3.2.8.

5.4 Results

5.5 CchH2 Homology Model

Prosa2003 analysis (see figure 5.9 A) of the best CchH2 model produced using the MODELLER alignment (see figure 5.3) identified four regions of the model structure which exhibited higher energy than equivalent regions of PheA (black line) as determined by the alignment. These regions of high energy unique to the CchH2 model (i.e. those not observed in PheA) are correlated with the location of a number of the insertions (numbered I1-I8 in figure 5.3 in the sequence alignment. Regions of gap insertion within the PheA:CchH2 alignment which correlate with regions of high energy within this CchH2 model are:

- I1 - A β-sheet shared by the two structures is misaligned.

- I3 - A gap inserted into the PheA sequence begins immediately after an α-helix. In CchH2 this α-helix is longer than the equivalent PheA α-helix.
- I4 - A gap inserted in the PheA sequence aligns with an α-helix in CchH2. The α-helix concerned is longer in CchH2 than the equivalent PheA α-helix.

As the sequence of CchH2 is longer than that of PheA (531 residues compared with 514 residues) some residues within the sequences that correspond, or are equivalent, in the sequence alignment do not appear to correspond in the Prosa2003 energy profile graph of PheA and CchH2 (figure 5.9) and instead appear shifted; for example in the energy graph the peak associated with gap insertion (I2) is observed $\sim$ residue 145 in PheA and $\sim$ residue 150 in CchH2. This is due to the five residue gap was inserted into the PheA sequence at position I1.

The region of structure between insertions I6 and I7 in the sequence alignment also displays a slightly higher energy than the equivalent region in PheA. The overall trend in this region is comparable, taking into the slight displacement of residue numbering in CchH2.

Alternate ways of accommodating these eight regions of insertion into the alignment were identified using a combination of automated alignment and fold prediction programs, secondary structure prediction and manual alignment adjustment, as described in Chapter 2. The alternate insertion options deemed to be viable options based on fulfilment of the criteria can be seen in the table in figures 5.4 (positions I1-I4) and 5.5 (positions I5-I8).

These various insertion options generated for each insertion region (I1-I8) using the automated methods will now be discussed.

All options generated for the location of the five residue gap insertion 1 (I1) in PheA using the automated methods were disregarded as either the insertion was located within a region of secondary structure of PheA or it produced a misalignment between regions of PheA and CchH2 secondary structure. Four viable I1 options were suggested by manually adjusting this region of the alignment.

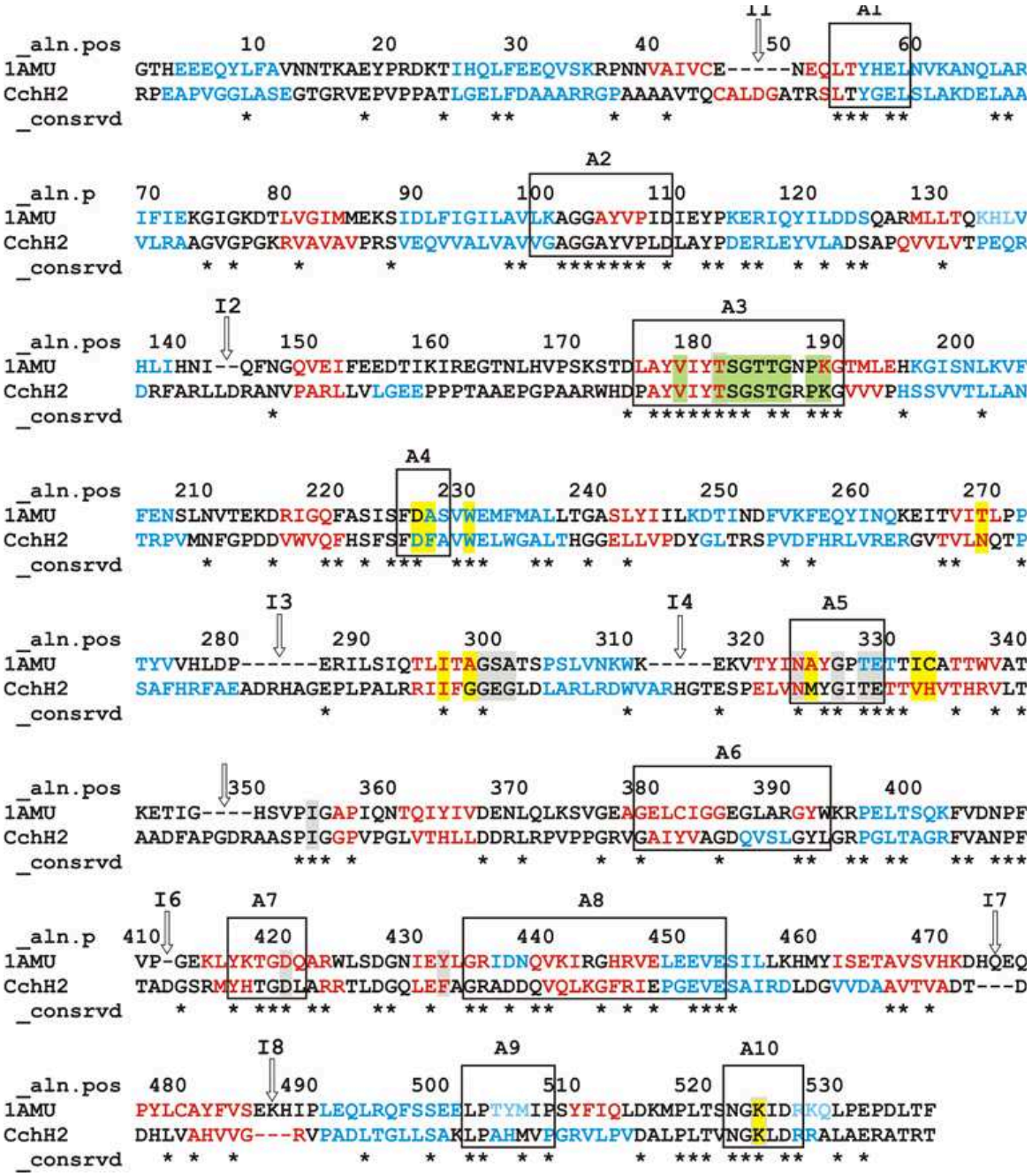


Figure 5.3: **MODELLER alignment of the GrsA A domain, PheA (pdb: 1AMU) and the module two A domain from the CchH NRPS (CchH2).** This alignment has 30.368% sequence identity. Residues are highlighted to illustrate the following: Green -AMP binding motif (Pfam: PF00501); Yellow - Substrate specific binding pocket residues (l-Thr); Grey - AMP binding pocket residues. Residues are coloured to represent the following: Red - β-sheet; Blue - α-helix; Light blue - 3-helix. Residues which form the conserved A domain motifs, determined by Marahiel *et al.*[22], are denoted using boxes and labelled by the motif code, i.e. A domain motif 1, A1. Similarly regions of insertion in the alignment are denoted using arrows and labelled by the insertion number, i.e. insertion 1, I1. Residues conserved within the alignment are denoted in the final line in the alignment using a $\star$.

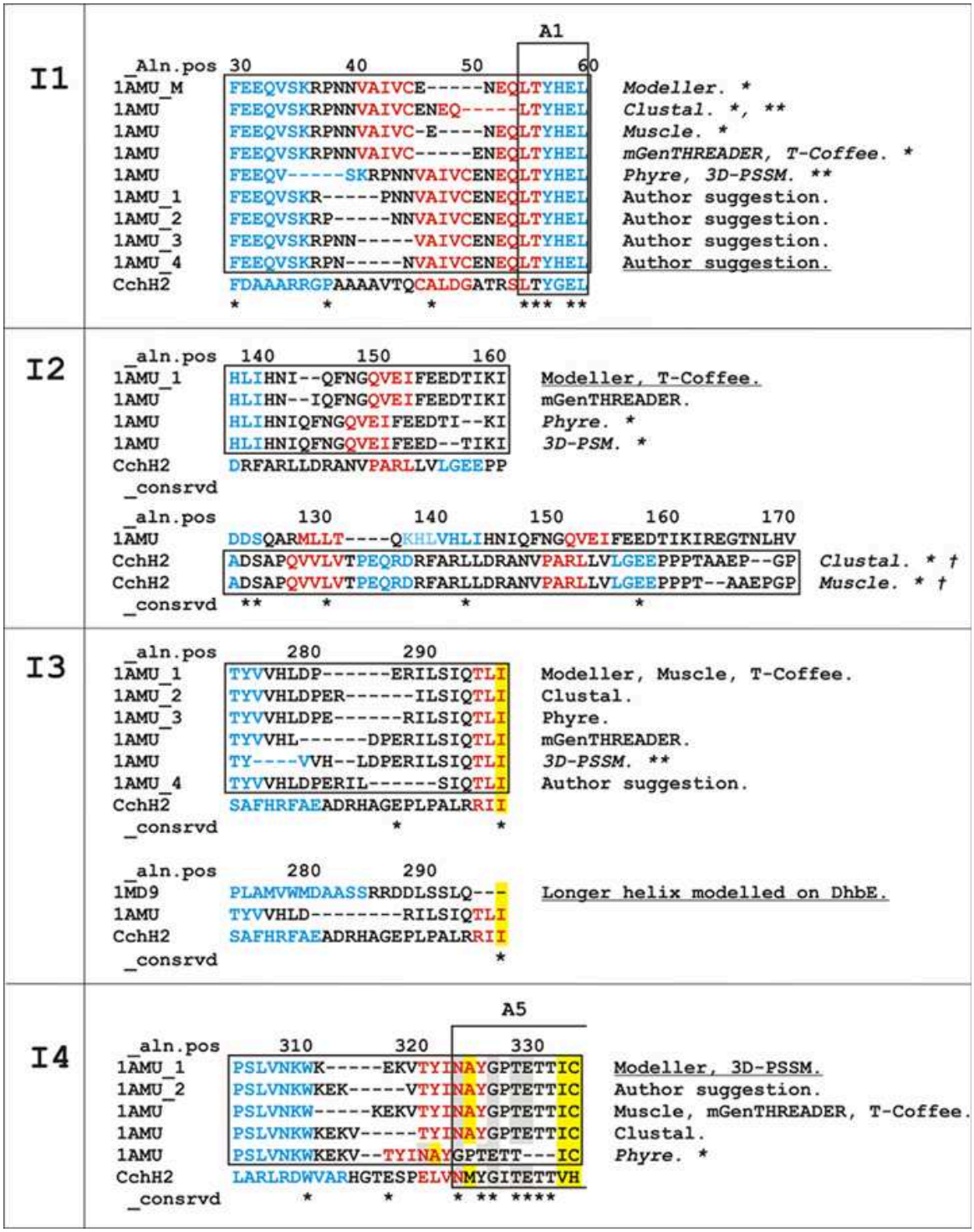


Figure 5.4: **Alternate alignment options for insertion locations one to four in PheA and CchH2 alignment.** Insertion variations were generated by automated alignment programs, fold prediction programs and manually. Viable insertion options are numbered next to the sequence identifier. The following symbols, found after the alignment method, denote the following: $\star$ - misalignment of secondary structure elements, $\star\star$ - insertion in secondary structure, † - distance violation as determined by MODELLER program. Residues are highlighted and coloured, and motifs denoted as in figure 5.3.

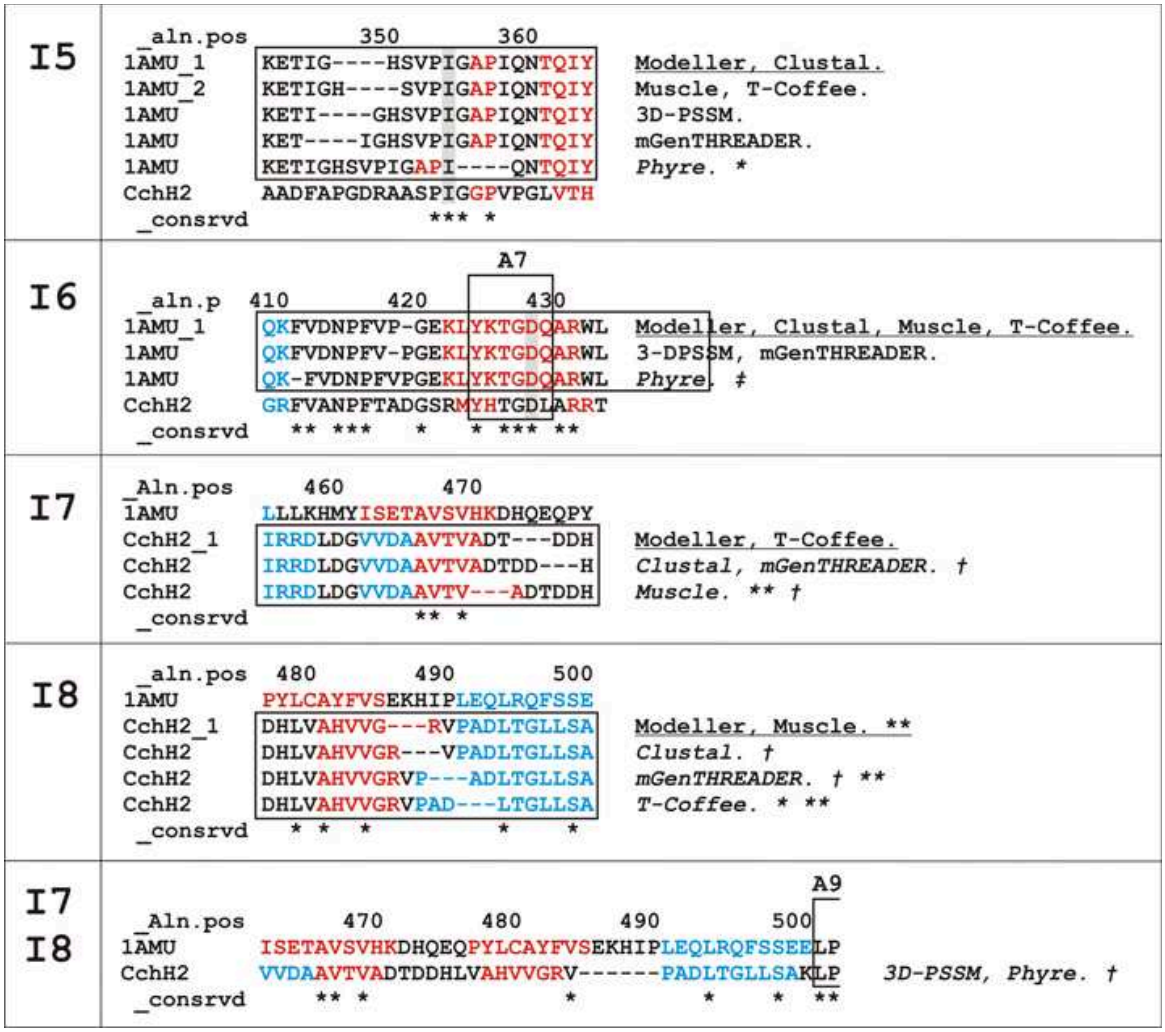


Figure 5.5: **Alternate alignment options for insertion locations five to eight in PheA and CchH2 alignment.**. Insertion variations were generated by automated alignment programs, fold prediction programs and manually. Viable insertion options are numbered next to the sequence identifier. The following symbols, found after the alignment method, denote the following: ⋆ - misalignment of secondary structure elements, ⋆⋆ - insertion in secondary structure, † - distance violation as determined by MODELLER program, ‡ - reduces number of conserved residues. Residues are highlighted and coloured, and motifs denoted as in figure 5.3.

The location of the two residue gap insertion 2 (I2) in the PheA sequence generated using the MODELLER alignment method (and also the T-Coffee method) was deemed to be the best of all the available I2 insertion options; analysis of the CchH2 model produced using the MODELLER alignment showed this region of structure in the model to be comparable in energy to that of PheA. The Clustal and Muscle programs suggested two alternate ways of aligning the sequence in this region, both very different from the ways suggested by the other methods. Both programs made two insertions in the general location of I2; one four residue gap insertion in the PheA sequence and one two residue gap insertion in the CchH2 sequence. These insertions were both discarded as they resulted in the misalignment of regions of equivalent secondary structure and violated the modelling distance criteria imposed by the MODELLER alignment.check routine.

Four viable insertion options were identified for the location of the six residue gap insertion I3. I3-1 suggested by MODELLER, Muscle and T-Coffee; I3-2 suggested by Clustal; I3-3 suggested by Phyre; and I3-4 produced by manual adjustment of the alignment. The insertion generated by mGenTHREADER was not selected as a viable insertion option as although it met the insertion viability criteria, the gap inserted into the PheA sequence aligns with a region of secondary structure in CchH2. The 3D-PSSM alignment split the six gap insertion in the PheA sequence at position I3 into two gap insertions; producing one gap two residues in length and one gap four residues in length. The first gap insertion was located in a region of secondary structure in PheA and therefore the 3D-PSSM insertions for this region were not treated as viable options.

Two viable alternate ways of inserting the five residue gap into the PheA sequence at position 4 (I4) were considered; I4-1 generated by MODELLER and 3D-PSSM and I4-2 a manually derived option. The further three options generated by: Muscle, mGenTHREADER and T-Coffee; Clustal; and Phyre; were all disregarded as they either reduced sequence identity or resulted in regions of secondary structure being aligned with gap insertions.

The two viable options for insertion position 5 (I5) where generated by MODELLER and Clustal, (I5-1) and Muscle and T-Coffee (I5-2). The Phyre I5 option violates the defined alignment criteria.

Only one insertion variation was used in position I6, that generated by MODELLER, Clustal, Muscle and T-Coffee (I6-1). As this region was modelled well using the initial MODELLER alignment, no alternate insertion option was investigated.

Examination of the options for the three residue gap insertion in the CchH2 sequence at position I7 identified only one viable option generated by programs MODELLER and T-Coffee. No viable insertion options were produced using MODELLER or the additional programs for the three residue gap insertion in the CchH2 sequence at position I8. An alternate position for the location of insertions 7 and 8 was suggested by programs Phyre and 3D-PSSM. This six position insertion, produced by combining the two three residue insertion at position I7 and I8, violates the modelling distance criteria imposed by the MODELLER alignment.check routine.

A full list of the various insertion options for each insertion position can be viewed in the alignment option tables (figures 5.4 and 5.5). Viable insertion options are numbered on the left hand side after the appropriate sequence tag and the various criteria violations denoted using a series of symbols.

The various options were combined into the original MODELLER alignment in an iterative fashion; as each insertion position was considered and a different insertion option combined into the alignment new models were built and the best model identified and assessed. Based on the results of the previous alignments model further alterations were made to the insertion regions at each stage until no further improvements in the models generated were observed. This process generated 13 new alignments on which 13 CchH2 models were built, assessed and the best selected for further analysis. The results of the assessments of these 13 new models can be viewed in the table in figure 5.6.

The various combinations of the alternate insertions positions that produced the best ten models can be seen in the table in figure 5.7.

The alignment 11 model was the highest ranking model in the Z-score and total energy analyses, having the lowest scores in both categories. The Prosa2003 Z-score and ga341 Z-score were -11.5 and -12.614 respectively and the total energy of the model, derived by

Alignment number	Best model	DOPE	Prosa Z-score	ga341 Z-score	Prosa Energy	ga341 Energy	Procheck* %	Procheck** %	Compactness	drms (nm)
1AMU	-	-6766.524	-12.75	-14.326	-421.30	-11.96	99.5	90.6	-	-
Modeller	56	-55746.609	-11.04	-12.156	-294.71	-7.619	98.6	92.4	0.257	0.042
1	70	-55067.031	-10.73	-12.016	-274.67	-7.219	98.2	92.4	0.270	0.032
2	14	-54605.340	-10.89	-11.6401	-285.09	-4.873	97.4	91.4	0.260	0.035
3	14	-55029.922	-10.67	-12.135	-270.56	-7.073	97.7	93.3	0.245	0.036
4	30	-57133.285	-10.65	-11.924	-269.50	-7.170	97.6	91.4	0.274	0.041
5	14	-55288.660	-10.32	-11.752	-261.26	-6.989	97.6	91.4	0.240	0.032
6	2	-55777.520	-10.93	-11.904	-287.56	-6.510	97.4	90.5	0.271	0.034
7	2	-55609.580	-10.65	-12.020	-269.44	-7.088	97.7	90.3	0.245	0.031
8	59	-55658.304	-10.93	-12.502	-287.50	-7.584	98.4	91.9	0.257	0.029
9	59	-55733.031	-10.96	-11.465	-289.21	-6.357	98.6	92.4	0.271	0.030
10	63	-55687.223	-10.85	-11.655	-282.11	-5.895	97.9	91.2	0.256	0.032
11	73	-56184.488	-11.05	-12.614	-294.75	-7.851	98.2	93.1	0.243	0.034
12	38	-55977.273	-10.48	-11.901	-258.53	-6.240	98.1	91.9	0.260	0.030
13	2	-55633.925	-10.69	-11.815	-272.11	-6.905	98.4	91.7	0.270	0.032
With DhbE	80	-56158.953	-10.93	-12.747	-287.54	-8.385	98.1	91.9	0.252	0.044
DhbE_loop	**18**	**-56322.180**	**-11.07**	**-12.606**	**-296.13**	**-8.405**	**98.4**	**92.4**	**0.254**	**0.044**

Figure 5.6: **Evaluation results of the best homology model,** as defined by the MODELLER Objective Function, produced by MODELLER from the 13 manually curated alignments, alignment 11 with one helix of DhbE and the model produced by the later alignment with additional loop modelling.

Alignment number	Sequence identity %	I1	I2	I3	I4	I5	I6	I7	I8
1	30.174	M	1	2	1	2	1	1	1
2	29.981	1	1	2	1	2	1	1	1
3	30.174	2	1	4	1	2	1	1	1
4	29.767	3	1	3	1	1	1	1	1
5	30.174	4	1	1	1	1	1	1	1
6	30.368	4	1	1	1	2	1	1	1
7	30.174	4	1	1	2	1	1	1	1
8	30.174	4	1	2	1	1	1	1	1
9	30.368	4	1	2	1	2	1	1	1
10	29.981	4	1	2	2	2	1	1	1
11	30.174	4	1	3	1	1	1	1	1
12	30.174	4	1	3	1	2	1	1	1
13	29.981	4	1	3	2	1	1	1	1

Figure 5.7: **Summary of PheA-CchH2 sequence alignments.** The insertion composition and percentage sequence identity of the 13 alignments generated during the iterative model refinement procedure.

ga341 and Prosa2003, was -294.75 and -7.851 respectively. This model was the second best performing in the DOPE and compactness analyses and in the Procheck analysis to identify the percentage of residues in the most favourable regions as defined by the Ramchandran plot (denoted Procheck ** in the table in figure 5.6). Additionally it came equal fifth in the Procheck analysis to identify the percentage of residues in the most favourable and additionally allowed regions as defined by the Ramchandran plot (denoted Procheck * in the table in figure 5.6). Although the RMSD value for this model (after superposition with the template structure PheA) calculated using the superpose routine within MODELLER was equal ninth lowest (out of 14), the value for this model 0.034 was equal to the mean 0.034 (standard deviation = 0.0038). No model produced by any of the other alignments ranked as consistently across the analyses.

In addition to these analyses, the alignment 11 model exhibited the most improvement in the regions of high energy structure exhibited by the MODELLER alignment model when assessed using Prosa2003. This improvement can be seen by comparing the graphs in figure 5.9; the most improved region in the profile of the alignment 11 model (figure 5.9 graph B) is highlighted using a red arrow.

Alignment 11 only differs from the MODELLER alignment in two regions, I1 and I3, yet the region the ChH2 model structure that exhibits a reduction in energy equates to residues 300–340, a region of the alignment which contains insertion 4, which has remained unaltered. The alternate gap insertions incorporated at positions I2 and I4 have the effect of reducing the overall energy of the model and reducing the region of high energy associated with 300–340.

Four regions of the CchH2 model structure were still of higher energy than the corresponding regions of PheA structure. The residues in these high energy regions of structure, and the secondary structures they form in CchH2, were identified as:

1. Residues 39-44 which form a loop.
2. Residues 272-280 which form an α-helix which is predicted to be longer in CchH2 than in PheA.

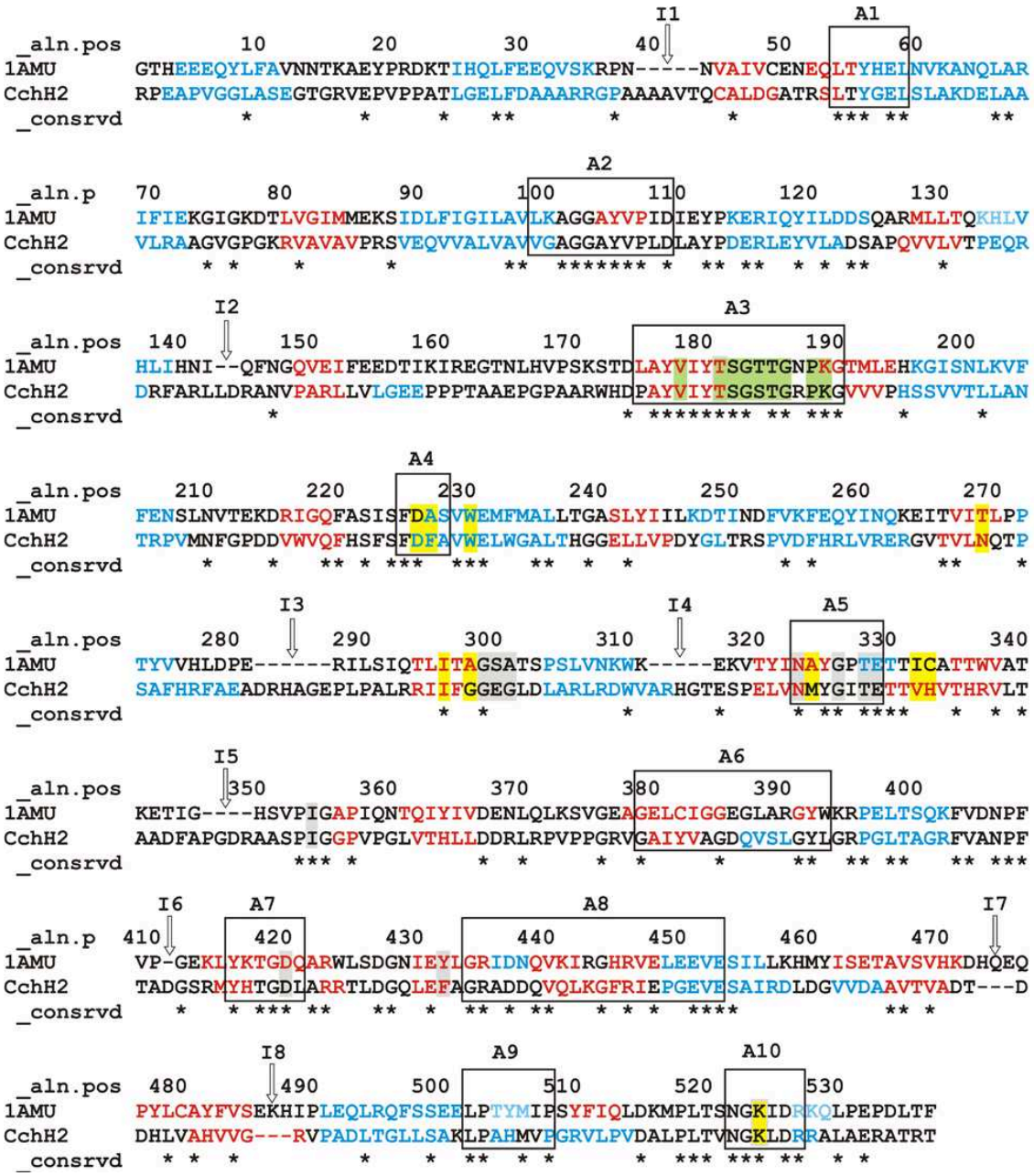


Figure 5.8: **Optimised PheA-CchH2 alignment.** Alignment 11 of the PheA A domain (1AMU) and the module two A domain from CchH NRPS (CchH2) produced during the iterative alignment refinement procedure. This alignment has 30.174% sequence identity. Residues are highlighted and coloured, motifs, insertions and conserved residues denoted as in figure 5.3.

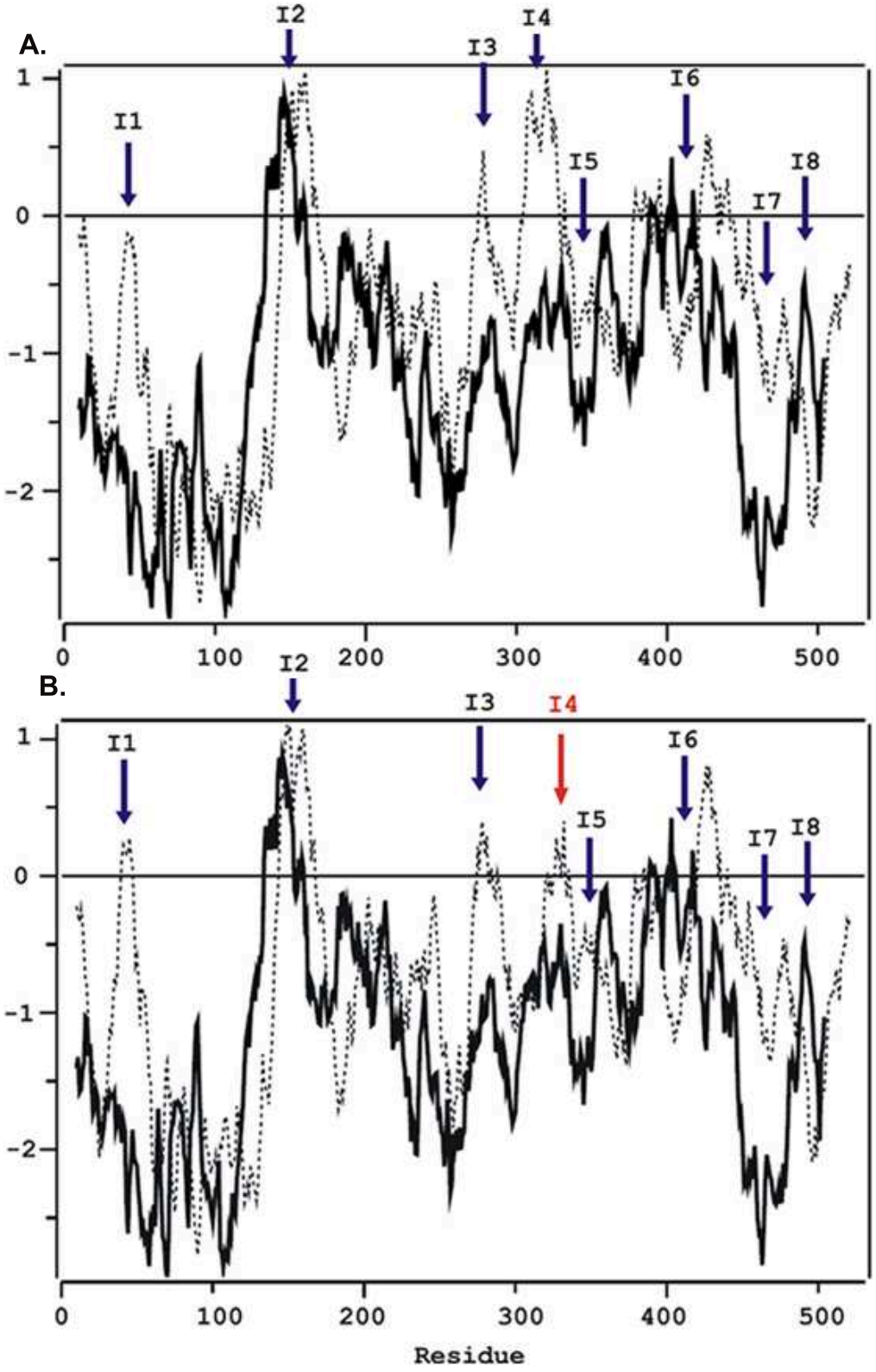


Figure 5.9: **Prosa2003 PheA and CchH2 energy profiles.** Graph showing the Prosa2003 energy profiles of: **A:** PheA (solid line) and the CchH2 homology model (broken line) built using the MODELLER alignment; and **B:** PheA (solid line) and the CchH2 homology model (broken line) built using alignment 11. The regions in the sequence alignment that contain the inserted gaps (I1–I8) are denoted using dark blue arrows and the red arrow in graph B indicates a region of improvement in the structure produced by the alignment refinement iteration which is indicated by a drop in energy.

3. Residues 330-340 which are highly conserved between PheA and CchH2. This region contains two substrate binding pocket residues which are associated with the functionality of the enzyme. One of the equivalent residues in PheA Ile 314 (pdb: 330) is an outlier in the Ramachandran plot, this is also true of the equivalent CchH2 residue Val 332.

4. Residues 426-429 which form a loop.

As residues 330–340 in CchH2 contain a residue associated with function which is an outlier on the Ramachandran map (as observed in PheA) no further refinement to this region was made. Instead the regions of structure comprising the longer CchH2 α-helix at positions 272–280 and the loops formed by residues 39–44 and 426–429 were selected for refinement.

Modelling of the longer CchH2 α-helix on the equivalent region of DhbE using the alignment in figure 5.10 greatly reduced the energy profile of this region of structure in the resulting best model as illustrated in graph A in figure 5.11. The inclusion of DhbE as a template for this region produced a model which scored worse than the alignment model 11 in the following areas: the Prosa2003 Z-score and energy, the DOPE score, the Procheck analyses, the compactness score and the RMSD value of CchH2 when compared to PheA. The Z-score and total energy calculated by GA341 for the "With DhbE" model showed an improvement on those calculated for the alignment 12 model. These results are presented in the table in figure 5.6 where this alignment is referred to as "With DhbE". The region of CchH2 structure modelled on DhbE template did show a marked drop in energy in the Prosa2003 energy profile graph (see figure 5.11 graph A). This Prosa2003 energy profile for this region of the CchH2 structure is now comparable with the equivalent region of PheA structure and as such this model was accepted.

Loop refinement of this model was subsequently performed. First the loop containing residues 39–44 was optimised and the best model, model 18, selected. Subsequently the loop comprising residues 436–439 was optimised. Analysis of the models produced after refinement of this region concluded that no significant improvements were made, either to

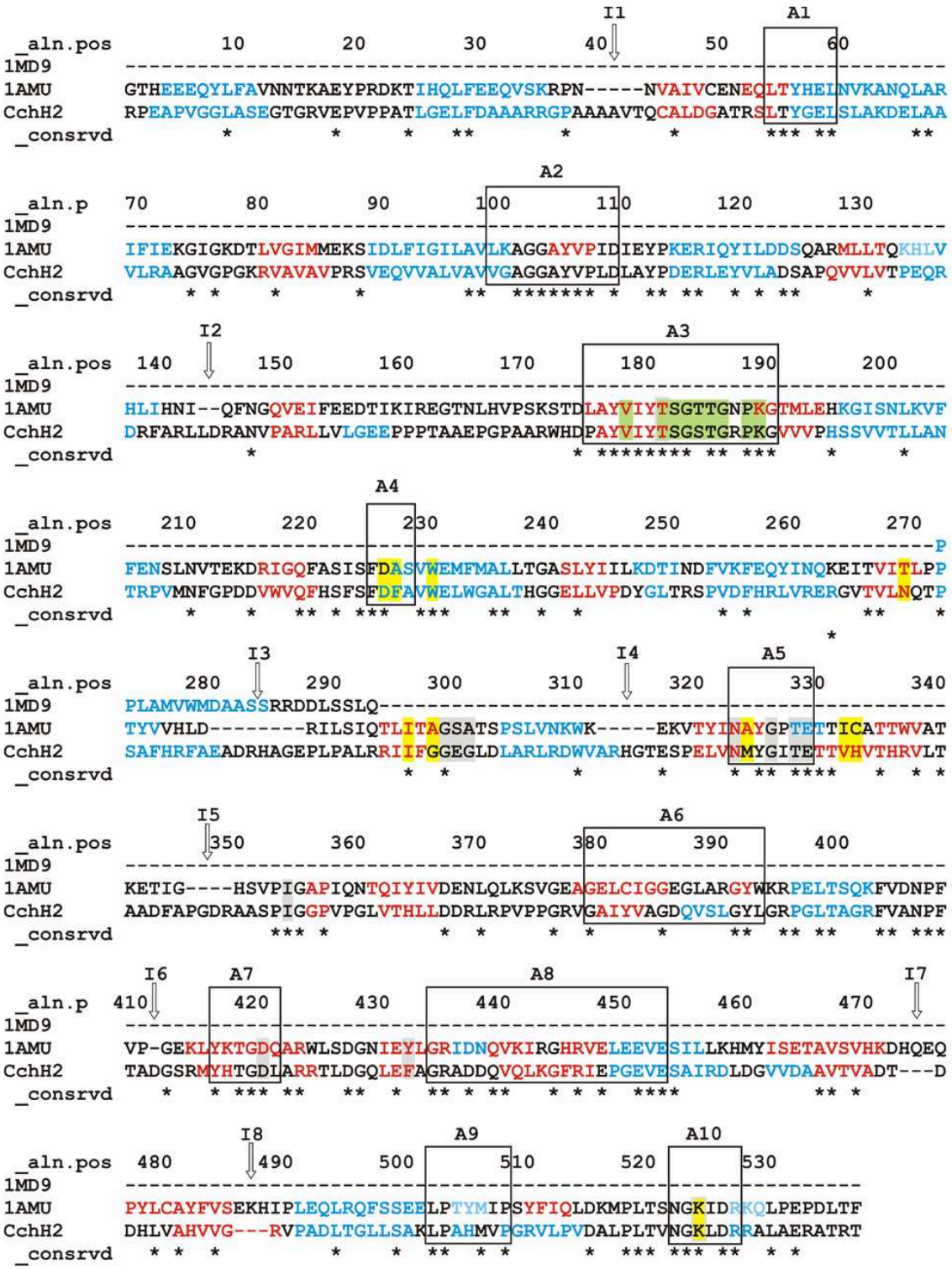


Figure 5.10: **Optimised PheA and DhbE-CchH2 alignment.** The alignment used to produce the CchH2 homology model. Alignment 11 from the iterative alignment refinement procedure, modified to model the residue 272–280 α-helix in CchH2 on the DhbE A domain equivalent helix. This alignment has 30.174% sequence identity. Residues are highlighted and coloured, and motifs denoted as in figure 5.3.

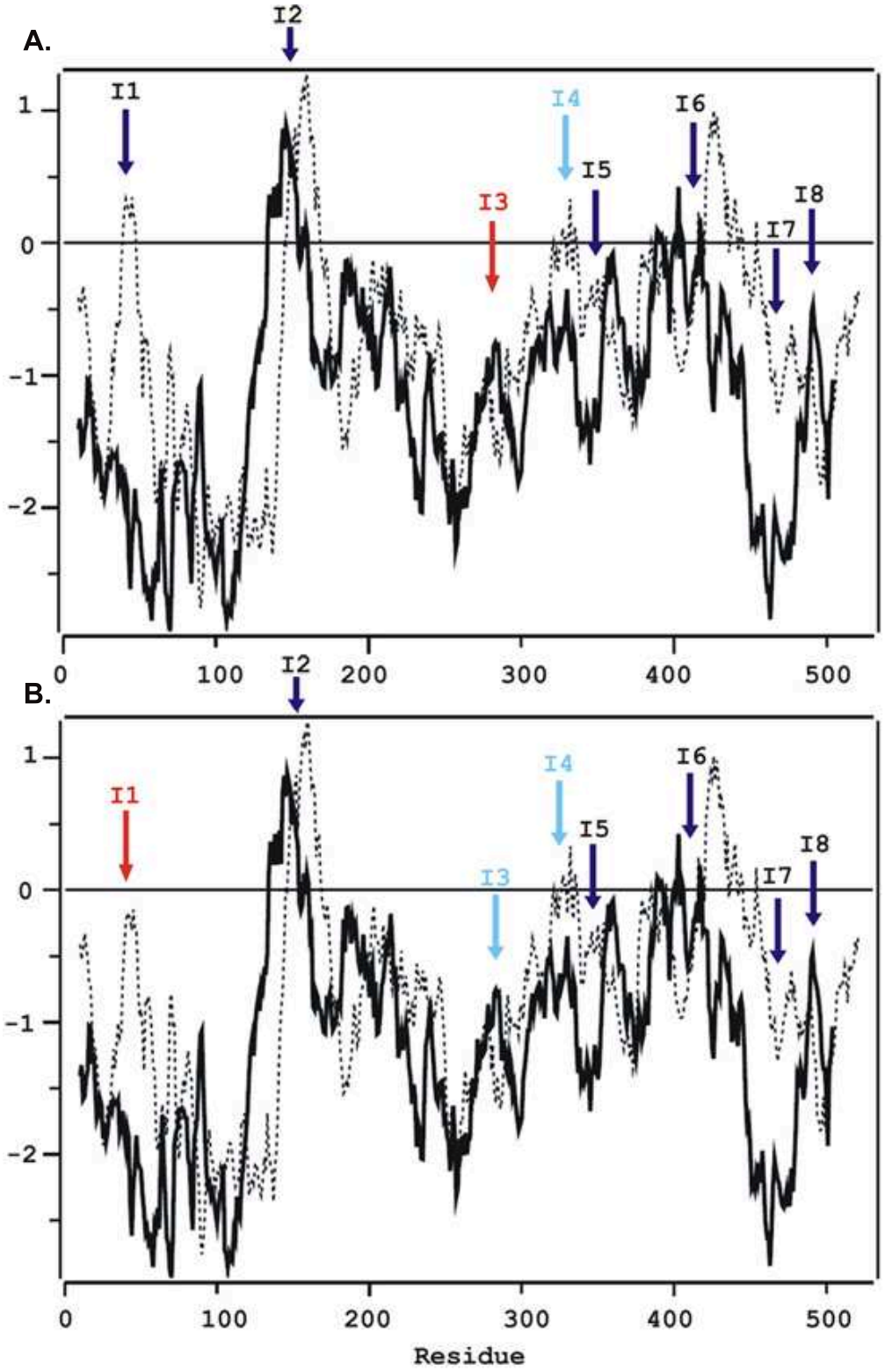


Figure 5.11: **Prosa2003 PheA, DhbE, PheA-CchH2, and Phe,DhbE-CchH2 model energy profiles.** Graph showing the Prosa2003 energy profiles of: **A:** PheA (solid line) and the CchH2 homology model (broken line) built using alignment 11 and templates PheA and DhbE; and **B:** PheA (solid line) and the CchH2 homology model (broken line) built using alignment 12, templates PheA and DhbE, and after the refinement of the loop comprising residues 39–44. The regions in the sequence alignment that contain the inserted gaps (I1-I8) are denoted using dark blue arrows, red arrows indicate regions of improvement obtained in this refinement iteration and light blue arrows indicate regions of improvement carried out at a previous stage.

this region of structure or the overall model. Model 18 was selected as the final CchH2 homology model. The results of the analyses of this model (alignment DhbE-loop) are in the table in figure 5.6. Comparison of the results for the models produced using the MODELLER alignment, alignment 11, With DhbE alignment and after refinement of loop 39–44 (DhbE-loop) shows that no one model is the highest ranking across all of the analyses. The DhbE-loop model has the lowest DOPE, Z-score (Prosa2003), and energies as calculated by Prosa2003 and GA341. The "With-DhbE" model has the lowest Z-score as calculated by GA341. The MODELLER model has the highest percentage of residues in the favourable and additionally allowed regions of the Ramachandran map; whereas the alignment 11 model has the highest percentage of residues in the favourable region of the Ramachandran map, the lowest compactness score and lowest RMSD when compared with the structure of PheA. The DhbE-loop model compactness score, 0.0254, is below the mean for all 16 models, 0.0258. The DhbE-loop model RMSD, 0.044, is however not within one standard deviation (0.005) of the mean RMSD value 0.035. The Prosa2003 energy profile for this model can be viewed (broken line), in comparison to that for PheA, in graph B of figure 5.11. The energy profile graph shows a drop in energy associated with the region of sequence improved by the loop refinement.

As no one model was the highest ranking across all the analyses and as in some analyses the results showed the modes were of comparable quality, the DhbE-loop model was chosen as the final model on which the MD simulations and docking studies would be performed.

5.6 Docking Results

Observations from the A domain structural data suggested and the results of simulations carried out on PheA in chapter 3 confirmed that the Asp 219 (CchH2:226) and Lys 501 (CchH2:518) residues at the top of the A domain binding pocket form electrostatic stabilising interactions with the substrate α-amino and α-carboxylate atoms. These residues are highly conserved and invariant, respectively, within the A domains. The existence of these interactions between the substrate and binding pocket and knowledge of the location of the

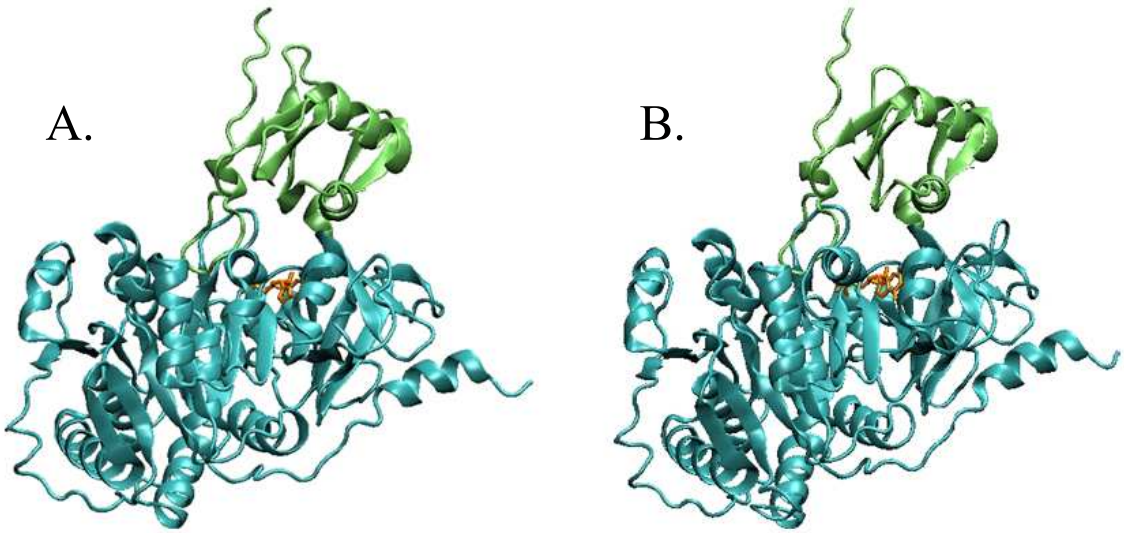


Figure 5.12: **Structure of final CchH2 homology model and PheA.** Image of **A:** PheA and **B:** the final CchH2 homology model. The A_{core} domain is coloured cyan, the A_{sub} domain green and the cofactors orange.

substrate binding pocket and the residues within it were used to help guide and assess the docking simulations.

Two initial attempts to dock the substrate ligand into various configurations of the CchH2 model were both unsuccessful when assessed using this criterion. The first, where the ligand was repeatedly docked into numerous CchH2 configurations taken from the CchH2-apo simulation, was unsuccessful as the relative positioning of binding pocket residues within the active site had evolved so that the substrate could not make the above outlined interactions with the enzyme. In the second attempt the CchH2 configuration to dock into was obtained by adding and minimising hydrogen atoms to the CchH2 homology model. The results from the docking runs using this CchH2 configuration did not position the substrate fully into the substrate binding pocket; instead the substrate was placed in close proximity to the AMP molecule and Mg^{2+}.

By comparing the CchH2 homology model and PheA crystallographic structures with the PheA structures from the end of the MD simulations in Chapter 3, a difference in the location of the Mg^{2+} was identified. Analysis of the MD simulations in Chapter 3, revealed a slight displacement of the positioning of the Mg^{2+} ion during the MD simulations. These simulations also revealed the Mg^{2+} coordination; six oxygen atoms two generally contributed by the AMP phosphate group, two by the carboxylate group of a glutamic acid

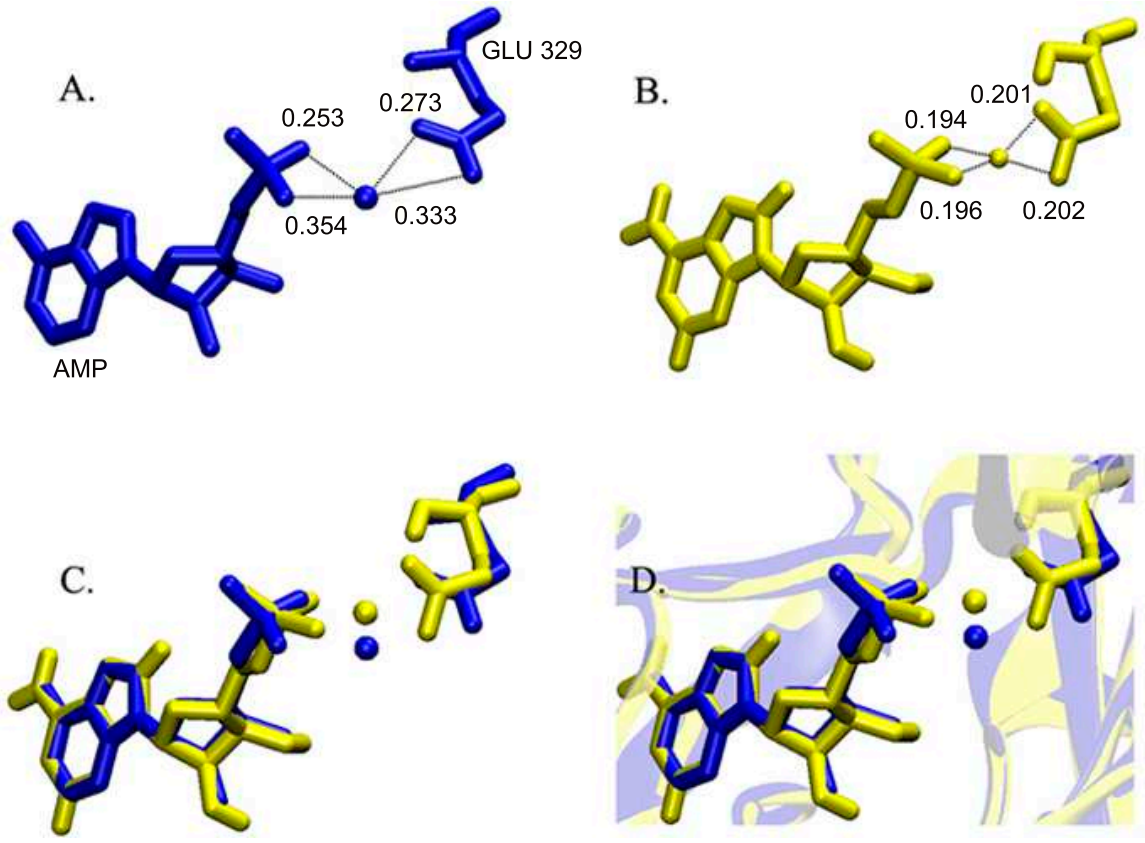


Figure 5.13: **CchH2 Magnesium ion positioning. A:** AMP, Mg^{2+} and Glu 329 positioning as in the CchH2 homology model. **B:** AMP, Mg^{2+} and Glu 329 positioning after minimisation in Gromacs. **C:** and **D:** AMP, Mg^{2+} and Glu 329 positioning pre- (blue) and post- (yellow) Gromacs minimisation.

residue sidechain, and two by water molecules. The Mg^{2+} ion was not observed during determination of the structure of the PheA enzyme and was instead placed into the crystal data file during the refinement stage. Energy minimisation of the CchH2 system (homology model and cofactors) - where all non-hydrogen atoms bar the Mg^{2+} ion were tethered - resulted in a slight displacement of the Mg^{2+} ion which was comparable to that seen in the PheA minimised and simulated structures. The displacement of the Mg^{2+} ion can be seen in figure 5.13. Using this CchH2 configuration for the docking simulations produced fruitful results, which will now be discussed for each substrate in turn.

Figure 5.14 shows the highest scoring docked L-Thr conformation from each of the ten clusters and the orientation of these structures in the CchH2 cofactor system. In figures A, B and C the ten docked conformations have been split into three groups on the basis of the orientation of the substrate α-amino and α-carboxylate atoms in relation to the Asp and Lys binding pocket residues. Group A contains highest ranking substrate conformations -

from clusters 1, 2 and 3. In this group the substrate appears twisted in the binding pocket; the substrate sidechain is pointing towards the Asp binding pocket residue, the substrate α-amino group towards the AMP phosphate group and the substrate α-carboxylate group towards the α-amino group of the Lys 518 enzyme residue. Group B contains substrate conformations from clusters 4, 8 and 9. In this group the substrate appears upside down in the binding pocket; the substrate sidechain is pointing towards the AMP phosphate group and the Lys binding pocket residue, the substrate α-carboxylate group is pointing down into the binding pocket (where the substrate sidechain is expected to be) and the substrate α-amino group is making the required interaction with the Asp binding pocket residue sidechain carboxylate group. Group C contains substrate conformations from clusters 5, 6, 7 and 10. In this group the substrate is oriented in the most productive conformation; the substrate sidechain is pointing down into the substrate binding pocket, the substrate α-carboxylate group is orientated towards the amino group of the sidechain of the Lys binding pocket residue and the substrate α-amino group is making the required interaction with the Asp binding pocket residue sidechain carboxylate group. From this group of docked substrate structures the highest ranking, lowest energy, conformation was selected - that from rank five. The orientation of this docked substrate is seen in figure 5.14 D. In this figure the distance between the binding pocket aspartic acid residue and substrate α-amino group atoms, and the binding pocket lysine residue and the substrate α-carboxylate atoms is displayed and the substrate binding pocket residues labelled. The L-Thr substrate hydroxyl sidechain group points towards the AMP molecule and the sidechain methyl group points down into the substrate binding pocket.

Figure 5.15 shows the highest scoring docked serine conformation from each of the ten clusters and the orientation of these structures in the CchH2 cofactor system. In figures A, B and C the ten docked conformations have been split into three groups on the basis of the orientation of the substrate α-amino and α-carboxylate atoms in relation to the Asp and Lys binding pocket residues. The orientation of the serine substrates in each of these groups follows the trend seen in L-Thr docking results. Group A contains highest ranking substrate conformations - from clusters 1, 2, 3, 5, 6, 9 and 10. In this group the serine substrate appears twisted in the binding pocket; the substrate sidechain is pointing towards the Asp

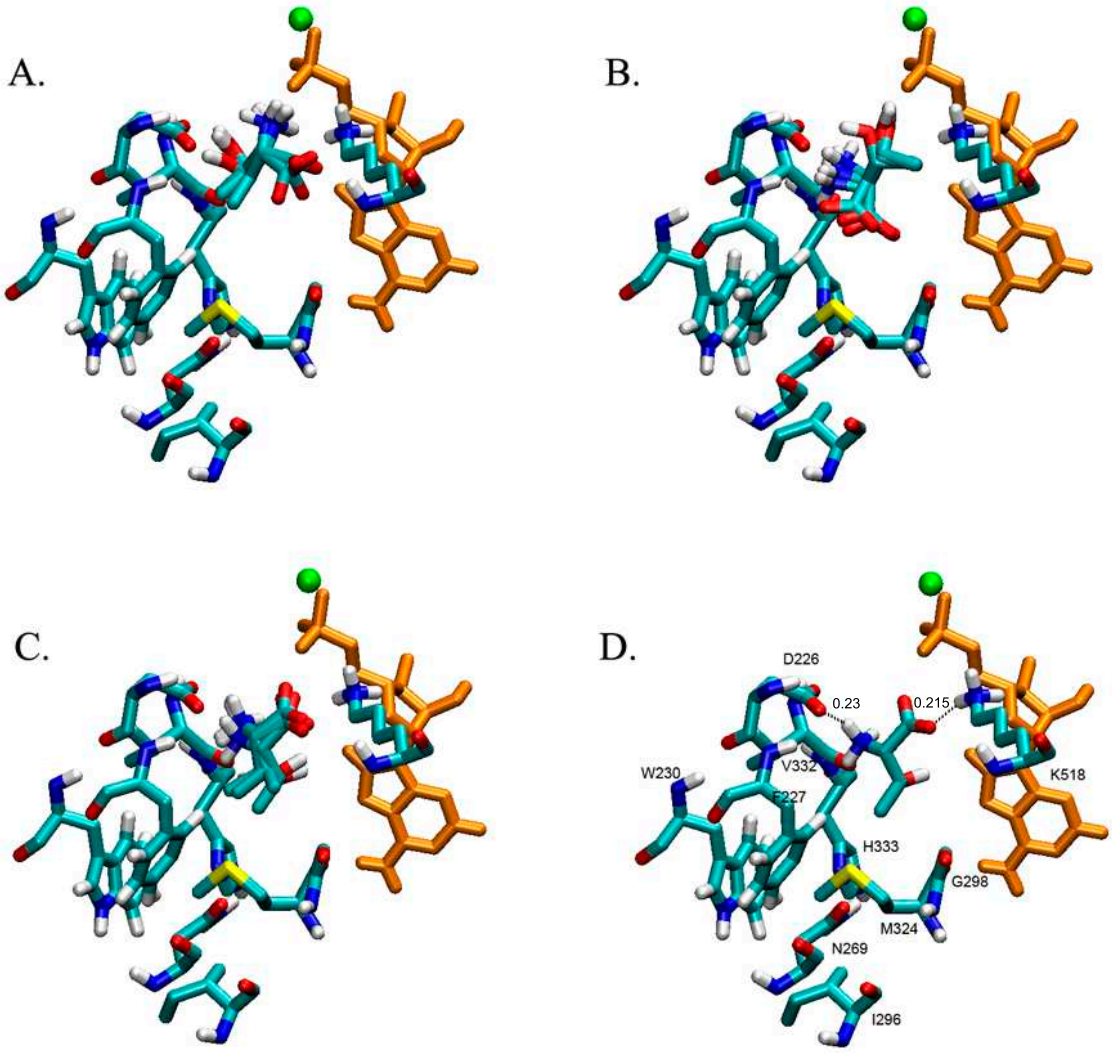


Figure 5.14: **Results of docking of substrate L-Threonine into CchH2.** The orientation of the docked substrate in the CchH2 cofactor system is shown. The highest ranking (lowest energy) threonine conformation in each of the ten clusters is shown. The substrate docked conformations are shown in groups according to the similarity in the orientation of the substrate α-amino and α-carboxylate atoms in relation to the Asp and Lys binding pocket residues. The AMP molecule is coloured orange and the Mg ion green. The protein and substrate atoms are coloured as follows; carbon atoms-cyan, oxygen-red, nitrogen-blue, sulphur-yellow and hydrogen-white. Only polar hydrogen atoms are displayed. A) Substrate conformations from clusters 1, 2 and 3. B) Substrate conformations from clusters 4, 8 and 9. C) Substrate conformations from clusters 5, 6, 7 and 10. D) The selected substrate conformer, cluster 5, and labelled binding pocket residues.

binding pocket residue, the substrate α-amino group towards the AMP phosphate group and the substrate α-carboxylate group towards the α-amino group of the Lys 518 binding pocket residue. Group B contains the substrate conformation from cluster 4. In this conformation the substrate appears upside down in the binding pocket; the substrate sidechain is pointing towards the AMP phosphate group and the Lys binding pocket residue, the substrate α-carboxylate group is pointing down into the binding pocket (where the substrate sidechain should be) and the substrate α-amino group is making the required interaction with the Asp binding pocket residue sidechain carboxylate group. Group C contains substrate conformations from clusters 7 and 8. In this group the substrate is oriented in the most productive conformation; the substrate sidechain is pointing towards the bottom of the substrate binding pocket, the substrate α-carboxylate group is orientated towards the amino group of the sidechain of the Lys binding pocket residue and the substrate α-amino group is making the required interaction with the Asp binding pocket residue sidechain carboxylate group. From this group of docked substrate structures the highest ranking, lowest energy, conformation was selected - that from rank seven.

The orientation of this docked substrate is seen in figure 5.15 D. In this figure the distance between the binding pocket aspartic acid residue and substrate α-amino group atoms, and the binding pocket lysine residue and the substrate α-carboxylate atoms is displayed and the substrate binding pocket residues labelled. The serine substrate hydroxyl sidechain group is pointing more towards the AMP molecule than towards the bottom of the substrate binding pocket.

Figure 5.16 shows the results of the docking calculation for the L-valine substrate and CchH2/cofactor system. Clustering analysis of the docked conformations identified only one orientation for L-valine within the CchH2 substrate binding pocket. The orientation of this valine ligand is well placed to form the required electrostatic interactions between the substrate α-amino and α-carboxylate atoms and binding pocket Asp and Lys residues respectively.

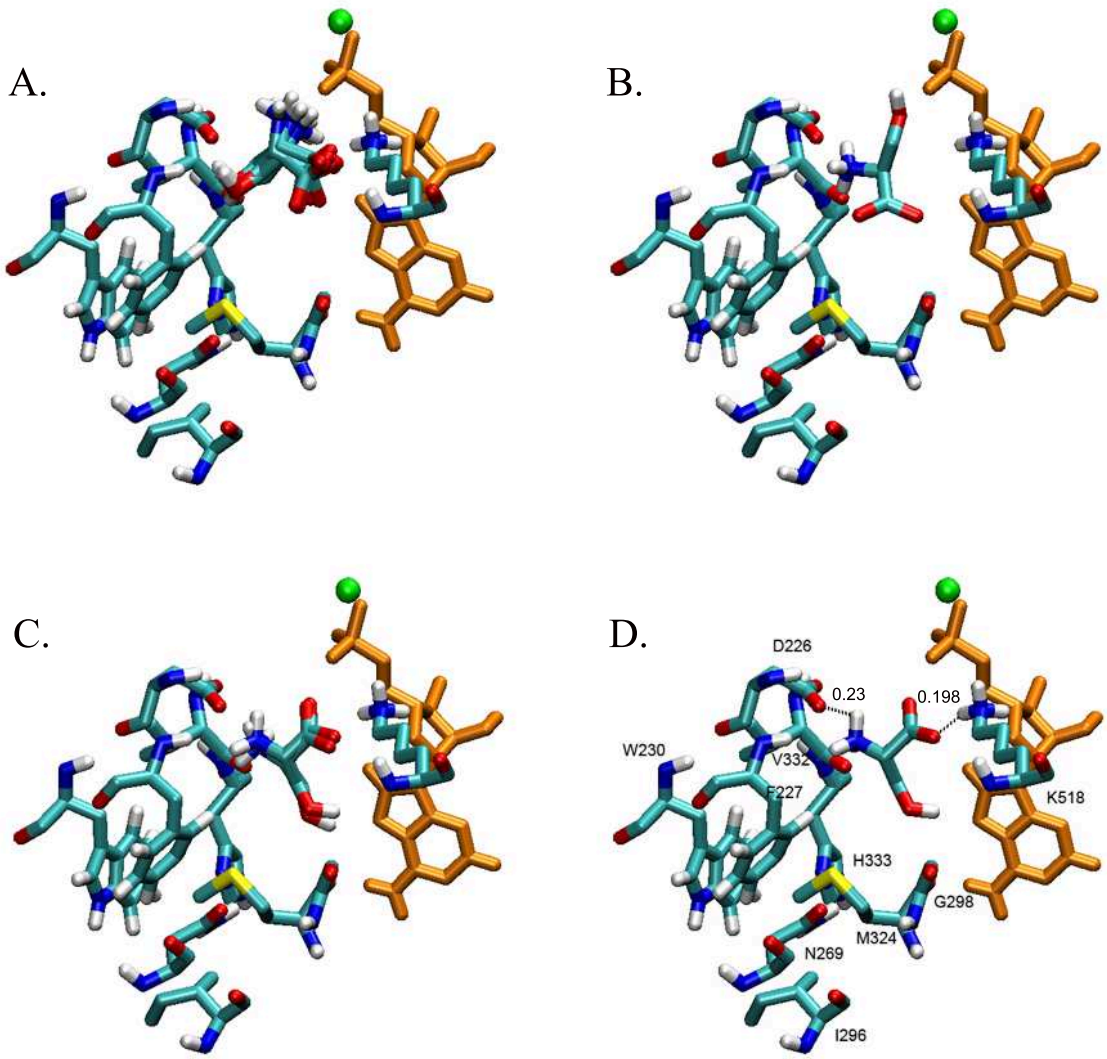


Figure 5.15: **Results of docking of substrate L-Serine into CchH2.** The orientation of the docked substrate in the CchH2 cofactor system is shown. The highest ranking (lowest energy) serine conformation in each of the ten clusters is shown. The substrate docked conformations are shown in groups according to the similarity in the orientation of the substrate α-amino and α-carboxylate atoms in relation to the Asp and Lys binding pocket residues. The AMP molecule is coloured orange and the Mg ion green. The protein and substrate atoms are coloured as follows; carbon atoms-cyan, oxygen-red, nitrogen-blue, sulphur-yellow and hydrogen-white. Only polar hydrogen atoms are displayed. A) Substrate conformations from clusters 1, 2, 3, 5, 6, 9 and 10. B) The substrate conformations from cluster 4. C) Substrate conformations from clusters 7 and 8. D) The selected substrate conformer, from cluster 7, and labelled binding pocket residues.

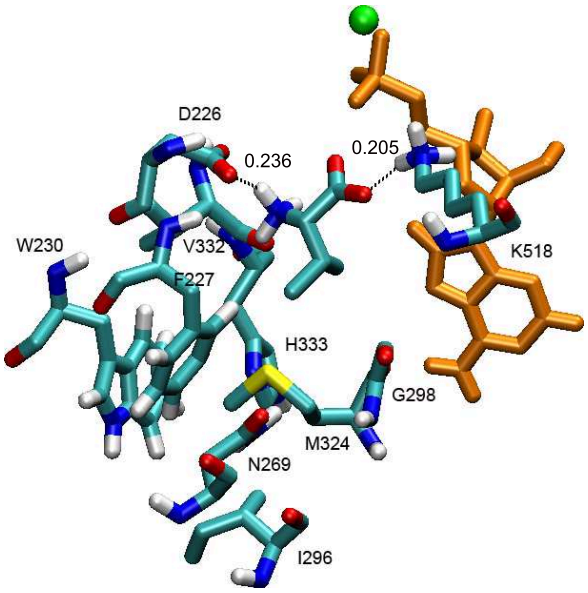


Figure 5.16: **Results of docking of substrate L-Valine into CchH2.** The only conformation for the orientation of the valine substrate in the CchH2 cofactor system is shown. The AMP molecule is coloured orange and the Mg ion green. The protein and substrate atoms are coloured as follows; carbon atoms-cyan, oxygen-red, nitrogen-blue, sulphur-yellow and hydrogen-white. Only polar hydrogen atoms are displayed.

5.6.1 Global Structural Stability

The relative conformational stability of CchH2 in each of the four simulations was assessed by calculating the RMSD of the Cα atoms from the initial structure (t=0) as a function of time. RMSD values for the Cα atoms from the large A_{core} and smaller A_{sub} domains were also obtained as shown in Chapter 3 when compared with the all atom Cα RMSD they may indicate interdomain motion. The RMSDs for these components of CchH2 in each of the simulations will now be considered in turn. The trends observed for these regions of CchH2 will be compared for each simulation and also with the equivalent PheA simulation RMSDs.

The RMSDs for these regions in the CchH2-apo state simulation can be seen in figure 5.17. The all Cα atom RMSD (black line) rises gradually to $\sim$0.35 nm by 2.5 ns, after 4 ns there it steadily increases to a final value of $\sim$0.4 nm. The A_{core} domain Cα atom RMSD (red line) displays very similar behaviour to that of the all Cα atom RMSD although the RMSD of this region is slightly lower overall. The A_{sub} domain Cα atom RMSD (blue line) shows much greater fluctuations throughout the simulation and is higher than that of the all Cα atoms. The RMSD for this region rises gradually to $\sim$0.35 nm by 2 ns, at 2.5 ns it begins

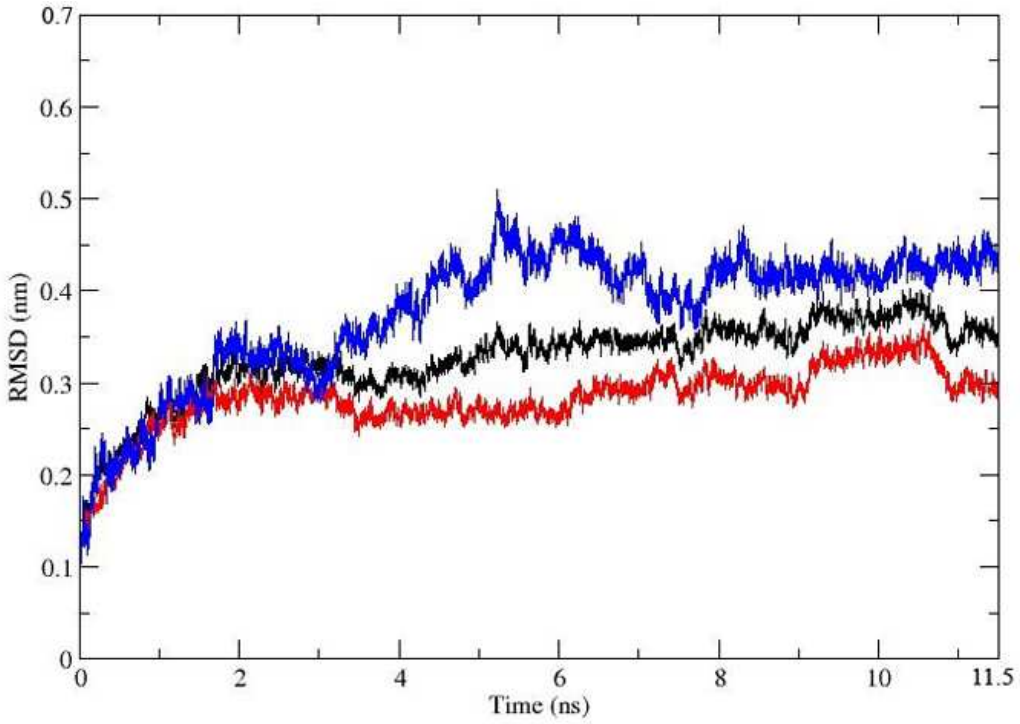


Figure 5.17: **RMSD CchH2-apo simulation.** Conformational drift of apo state CchH2, measured as C-α atom root mean square deviation (RMSD) from the starting structure. RMSDs vs. time are shown for the entire protein (black), the A_{core} domain (red) and the A_{sub} domain (blue).

to increase reaching a peak of 0.5 nm by 5 ns, after which time it reaches a relatively stable plateau of 0.45 nm.

The RMSD of the A_{core} domain in the CchH2 apo simulation shows this domain to be less stable than that of PheA in the PheA1-apo simulation. The A_{sub} domain exhibits greater structural drift in the CchH2 apo simulation than the PheA1-apo simulation. The all Cα atom RMSD does not indicate any relative A_{core} / A_{sub} domain rotation in the CchH2 apo simulation.

The RMSDs for all the Cα atoms (black), the A_{core} domain Cα atoms (red) and the A_{sub} domain Cα atoms (blue) for the CchH2-Thr simulation can be seen in figure 5.18. The all Cα atom RMSD rises gradually to a peak of $\sim$0.55 nm by 4.5 ns, after this time the RMSD declines to $\sim$0.5 nm remaining at this value until the end of the simulation except for a small peak at 10 ns where the RMSD briefly rises again to $\sim$0.55 nm. The A_{core} domain Cα atom RMSD rises to a plateau of $\sim$0.325 nm at 1.5 ns and remains constant for the rest of the simulation; this domain shows similar stability in the CchH2-Thr simulation as

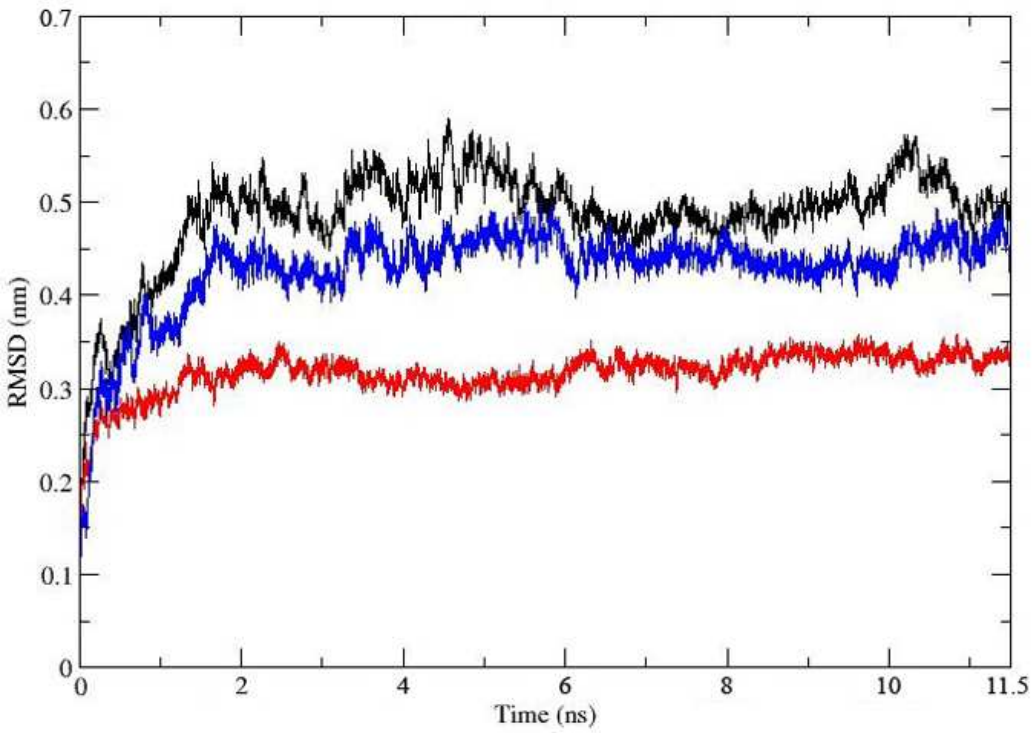


Figure 5.18: **RMSD CchH2-Thr simulation.** Conformational drift of apo state CchH2, measured as C-α atom root mean square deviation (RMSD) from the starting structure. RMSDs vs. time are shown for the entire protein (black), the A_{core} domain (red) and the A_{sub} domain (blue).

the CchH2 apo simulation. The A_{sub} domain Cα atom RMSD is similar to that of the all Cα atom RMSD although the RMSD is $\sim$0.5 nm lower overall. The A_{sub} domain exhibits greater structural drift than the A_{core} domain. The higher all Cα atom RMSD indicates there may be some relative A_{core} / A_{sub} domain motion in the CchH2-Thr simulation.

The RMSDs for all the Cα atoms (black), the N-terminal domain (A_{core} domain) Cα atoms (red) and the C-terminal domain Cα atoms (blue) for the CchH2-Ser simulation can be seen in figure 5.19. The all Cα atom RMSD rises to peak at $\sim$0.6 nm at $\sim$2 ns, after which it fluctuates between $\sim$0.55 nm and $\sim$0.45 nm for the rest of the simulation. Once again the A_{core} domain Cα atom RMSD indicates this domain to be the most stable on the timescale of the simulation; the RMSD rises to a plateau of $\sim$0.3 nm at 2 ns and remains constant until 6 ns where it rises again to reach a plateau of $\sim$0.325 nm. The A_{sub} domain Cα atom RMSD rises to $\sim$0.4 nm at 1ns which increase to $\sim$0.55 nm at 3 where it remains until $\sim$5 ns at which time it decreases to $\sim$0.4 nm and it fluctuates around this value until the end of the simulation.

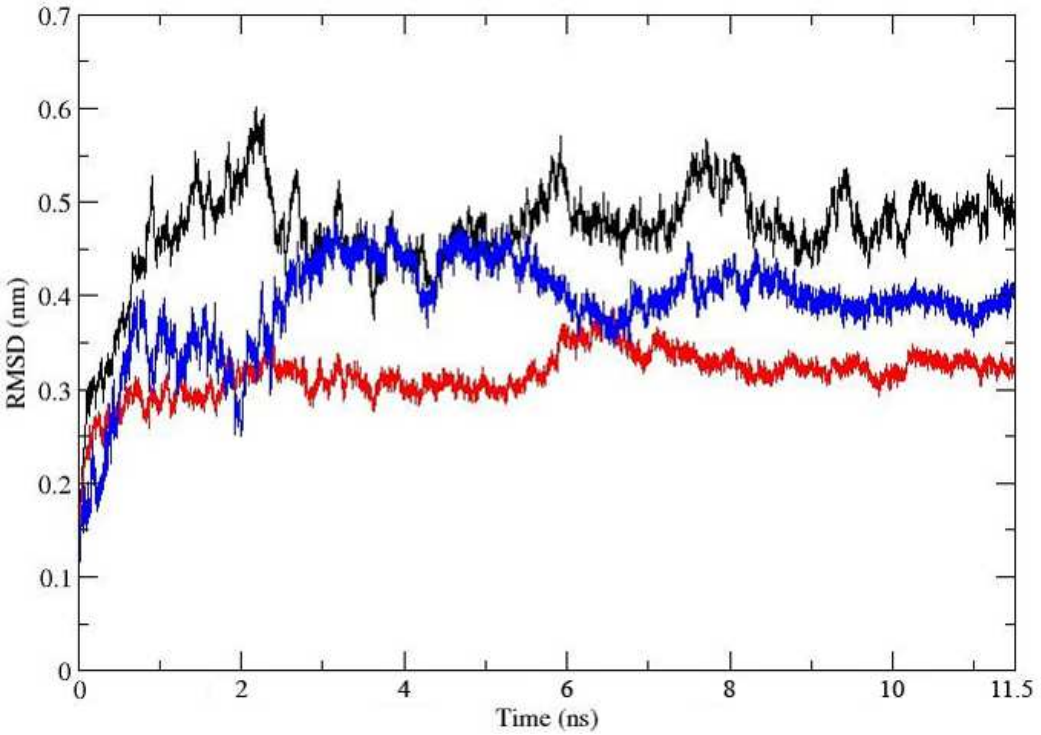


Figure 5.19: **RMSD CchH2-Ser simulation.** Conformational drift of apo state CchH2, measured as C-α atom root mean square deviation (RMSD) from the starting structure. RMSDs vs. time are shown for the entire protein (black), the A_{core} domain (red) and the A_{sub} domain (blue).

An average the A_{sub} domain is slightly more stable in the CchH2-Ser simulation than the CchH2 threonine simulation. A peak of the A_{sub} domain RMSD is observed between the second and sixth nanoseconds of the simulation. Overall the all Cα atom RMSD of CchH2 is higher than the RMSDs of the two domains indicating there may be some A_{core} / A_{sub} domain motion occurring in this simulation.

The RMSDs for all the Cα atoms (black), the A_{core} domain Cα atoms (red) and the A_{sub} domain Cα atoms (blue) for the CchH2-Val simulation can be seen in figure 5.20. The all Cα atom RMSD rises to plateau at $\sim$0.5 nm at $\sim$2 ns. The A_{core} domain Cα atom RMSD reaches a plateau of $\sim$0.4 nm at 4.5 ns. The A_{sub} domain Cα atom RMSD gradually rises to plateau at $\sim$0.45 nm at 7 ns. The A_{core} and A_{sub} domains exhibit greater structural drift in this simulation than in either the CchH2 holo Thr and Ser simulations. The A_{core} domain is less stable in this simulation with the RMSD for this region following a similar evolution to that of the A_{core} domain. The all Cα atom RMSD is higher than that of either the A_{core} or A_{sub} domain indicating there may be some small scale A_{core} / A_{sub} domain

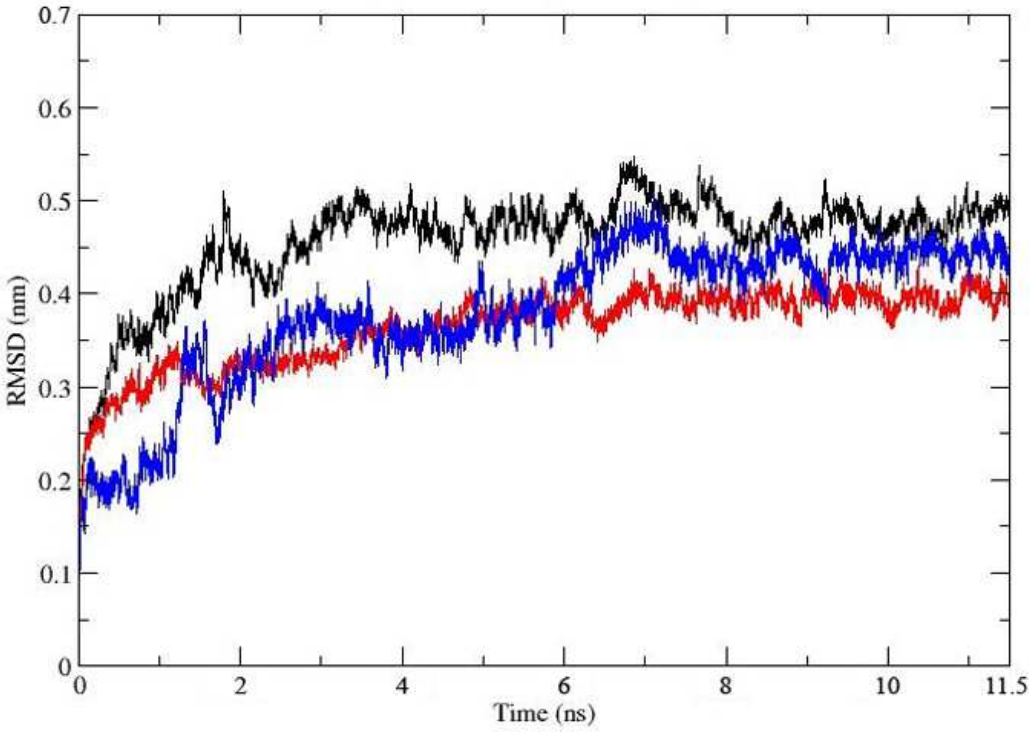


Figure 5.20: **RMSD CchH2-Val simulation.** Conformational drift of apo state CchH2, measured as C-α atom root mean square deviation (RMSD) from the starting structure. RMSDs vs. time are shown for the entire protein (black), the A_{core} domain (red) and the A_{sub} domain (blue).

motion.

RMSD analysis of the all, A_{core} domain and A_{sub} domain Cα atoms of the PheA protein in the simulations carried out in Chapter 3 identified a clear trend (large fluctuations in the all Cα atom RMSD and stable RMSD for the A_{core} domain and stable yet greater structural drift of the A_{sub} domain Cα atoms) which was later attributed to interdomain motion within the protein. This trend observed in some of the holo-state CchH2 simulations it is not observed in the apo-state CchH2 simulations. The large external loops of a protein often exhibit the greatest flexibility. This may be especially true when considering simulations carried out using a homology model; where low energy loops but not necessarily the correct loop has been obtained using theoretical methods. To exclude these regions from potentially contributing large fluctuations to the RMSD values, the all-atom Cα, A_{core} domain Cα and A_{sub} domain Cα regions of protein were decomposed to include only atoms withinα-helices and β-strands. The first α-helix in CchH2 (H1), which is very short, is considered to fall into the linker region, and also shows varying stability over time in both

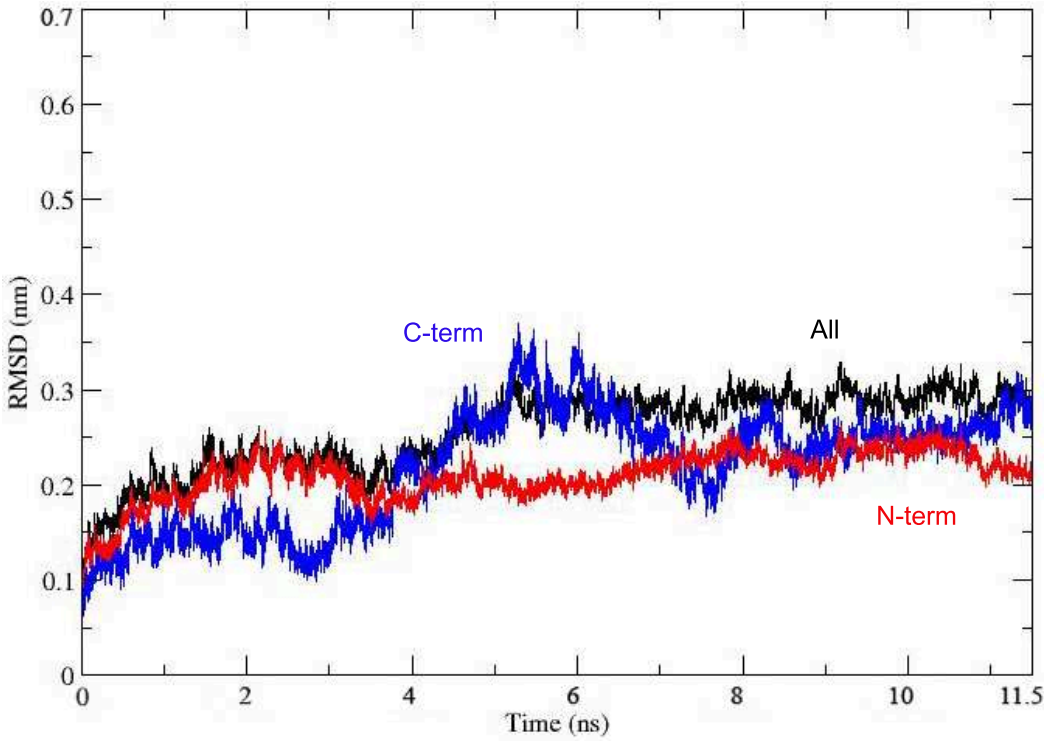


Figure 5.21: **RMSD N-terminal and C-terminal domain; CchH2-apo simulation.** Conformational drift of the combined structural components (α-helices and β-strands) of the protein in apo state CchH2, measured as C-α atom root mean square deviation (RMSD) from the starting structure. RMSDs vs. time are shown for both domains (black), the N-terminal domain (A_{core} domain) (red) and the C-terminal domain (A_{sub} domain) (blue).

the PheA and CchH2 simulations (DSSP data), was excluded from this analysis. These RMSDs calculated for each simulation will now be discussed. Analysis of the α-helices and β-strands of the all-atom Cα (black), A_{core} (red) and A_{sub} domain Cα (blue) atoms from the CchH2-apo simulation showed little deviation in the overall trend in the RMSDs of these regions, see figure 5.21, although the values have slightly decreased. To investigate the large RMSD fluctuations observed in the A_{sub} domain (C-terminal domain) Cα atom RMSD further the RMSD for this region was decomposed into the RMSDs for the α-helices (blue) and β-strands (red), (see in figure 5.22). This graph shows that on average the α-helices of the A_{sub} domain contribute more than the β-strands to the structural variation seen in the A_{sub} domain of the CchH2-apo state protein.

In the CchH2-Thr simulation (see figure 5.23), the all Cα atom RMSD rises to peak at $\sim$0.45 nm at $\sim$2 ns, after which it fluctuates between $\sim$0.5 nm and $\sim$0.35 nm for the rest of the simulation. The N-terminal domain (A_{core} domain) Cα atom RMSD rises to a plateau of $\sim$0.2 nm at 0.5 ns. The C-terminal domain (A_{sub} domain) Cα atom RMSD rises to

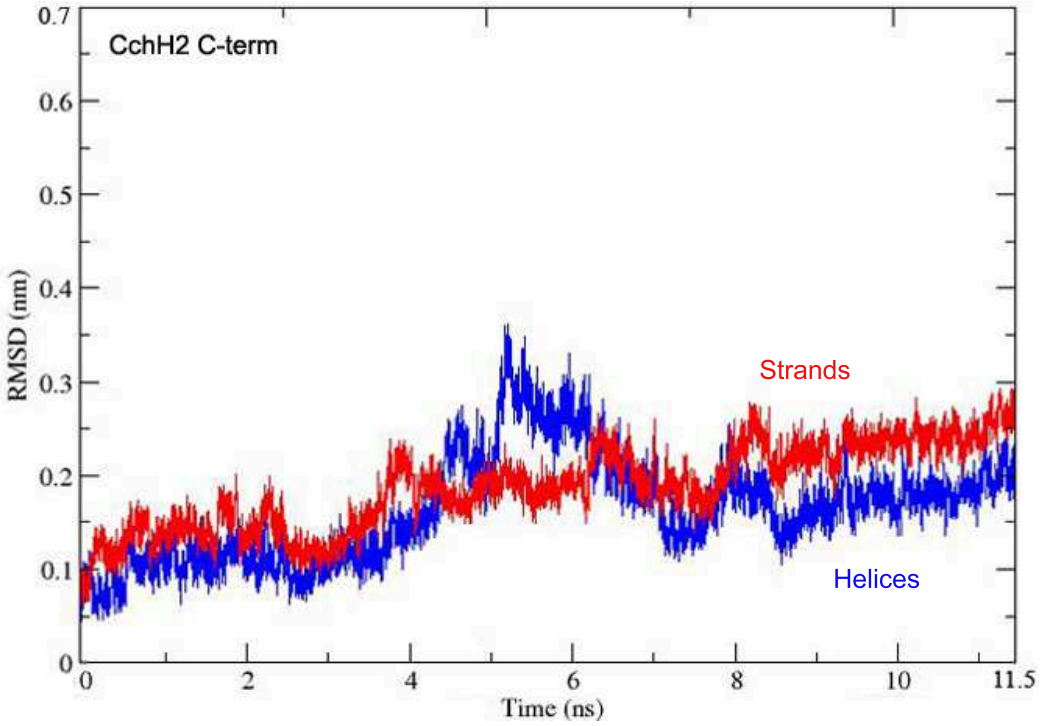


Figure 5.22: **RMSD secondary structure C-terminal domain; CchH2-apo simulation.** Conformational drift of the structural components (α-helices or β-strands) of the C-terminal domain (A_{sub} domain) in apo state CchH2, measured as C-α atom root mean square deviation (RMSD) from the starting structure. RMSDs vs. time are shown for the C-terminal domain α-helices (blue) and the C-terminal domain β-strands (red).

$\sim$0.15 nm at 0.2 ns, dropping to $\sim$0.1 nm until 4 ns where it begins to increase until 7 ns where it rises to plateau at $\sim$0.2 nm. By looking at just the helices and sheets of the domains, the A_{core} domain is shown to be less stable on the timescale of the simulation than the A_{sub} domain, which is different to what is observed in the PheA-holo simulations. The difference between the RMSD for the individual domains does not account for that of the overall Cα atom RMSD indicating there may be interdomain motion in this simulation.

In the CchH2-Ser simulation (see figure 5.24) the all Cα atom RMSD rises to peak at $\sim$0.55 nm at $\sim$2 ns, after which it fluctuates between $\sim$0.5 nm and $\sim$0.35 nm for the rest of the simulation. The N-terminal domain (A_{core} domain) Cα atom RMSD rises to a plateau of $\sim$0.25 nm at 0.5 ns. The C-terminal domain (A_{sub} domain) Cα atom RMSD rises to $\sim$0.25 nm at 5 ns, dropping to plateau at $\sim$0.2 nm at 6 ns. As in the CchH2 threonine simulation the helices and sheets of the A_{core} domain are less stable that those from the A_{sub} domain. The all Cα atom RMSD indicates that interdomain motion may occur in this simulation.

The RMSDs of the CchH2-Val simulation (see figure 5.25) display a different trend to

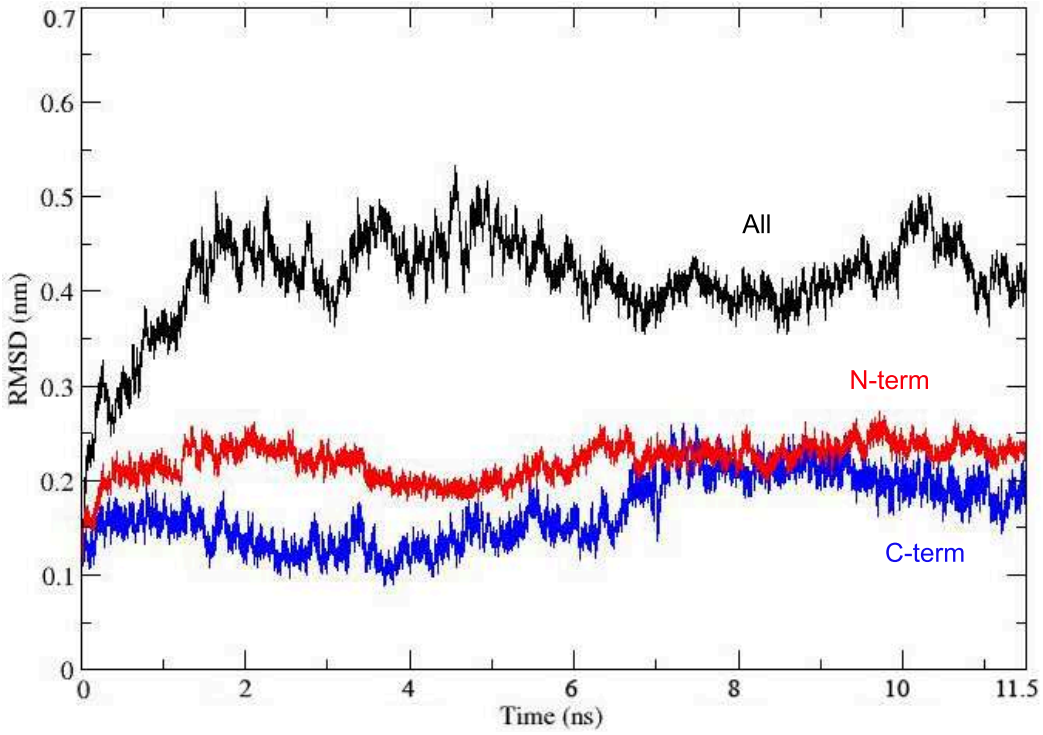


Figure 5.23: **RMSD N-terminal and C-terminal domain; CchH2-Thr simulation.** Conformational drift of the combined structural components (α-helices and β-strands) of the protein in apo state CchH2, measured as C-α atom root mean square deviation (RMSD) from the starting structure. RMSDs vs. time are shown for both domains (black), the N-terminal domain (A_{core} domain) (red) and the C-terminal domain (A_{sub} domain) (blue).

those observed in the threonine and serine simulations. In the CchH2-Val simulation, see figure 5.25, the all Cα atom RMSD rises to peak at $\sim$0.4 nm at $\sim$2 ns, after which it fluctuates between $\sim$0.4 nm and $\sim$0.33 nm for the rest of the simulation. The N-terminal domain (A_{core} domain) Cα atom RMSD gradually rises throughout the simulation starting at $\sim$0.25 nm at 1 ns and ending at $\sim$0.3 nm. The C-terminal domain (A_{sub} domain) Cα atom RMSD gradually rises to peak at $\sim$0.25 nm at 7.5 ns, dropping to end the simulation at $\sim$0.2 nm. As for the threonine and serine simulations the RMSD of the helices and sheets of the A_{core} domain indicates greater structural drift in this domain than the A_{sub} domain. The all Cα atom helices and sheets RMSD is higher than those of the individual domains however the difference is not as great as that observed in the threonine and serine simulations.

Overall, analysis of the RMSD of the various regions of CchH2 from the starting structure in each simulation demonstrated a high degree of structural stability over the simulation timescale. In the apo CchH2 simulation higher structural drift was exhibited by the A_{sub}

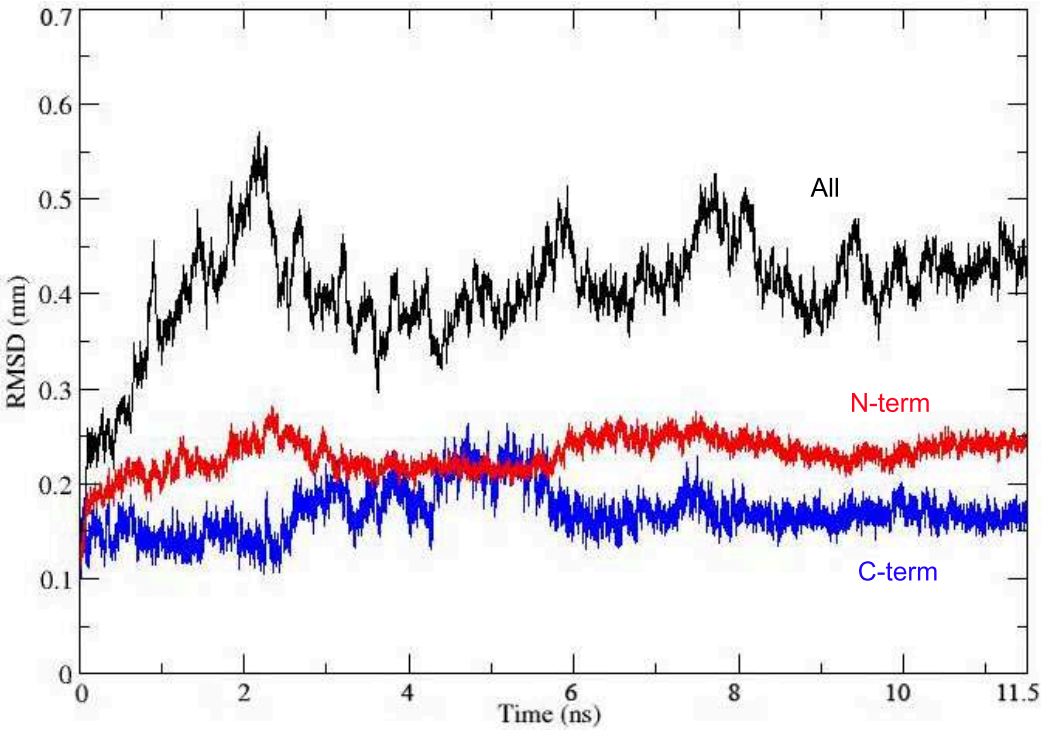


Figure 5.24: **RMSD N-terminal and C-terminal domain; CchH2-Ser simulation.** Conformational drift of the combined structural components (α-helices and β-strands) of the protein in apo state CchH2, measured as C-α atom root mean square deviation (RMSD) from the starting structure. RMSDs vs. time are shown for both domains (black), the N-terminal domain (A_{core} domain) (red) and the C-terminal domain (A_{sub} domain) (blue).

domain, which can largely be attributed to the α-helices, than by the whole protein or A_{core} domain. However the final RMSD values for each of these regions are comparable when the RMSDs of only the α-helices and β-sheets are analysed (figure 5.21).

In the holo CchH2 threonine and serine simulations, figures 5.23 and 5.24 respectively, higher structural drift was exhibited over all the C-α atoms than by the individual domains, suggesting motion of these domains relative to one another, as observed in the PheA simulations in Chapter 3. Decomposing the individual domains to look at the α-helices and β-strands revealed in each simulation the A_{core} domain exhibited greater structural drift than the A_{sub} domain. In the holo CchH2 valine simulation, figures 5.25 the most structural drift is observed exhibited over all the C-α atoms, although the A_{core} and A_{sub} domain C-α atoms exhibit a similar but slightly lesser degree of structural drift.

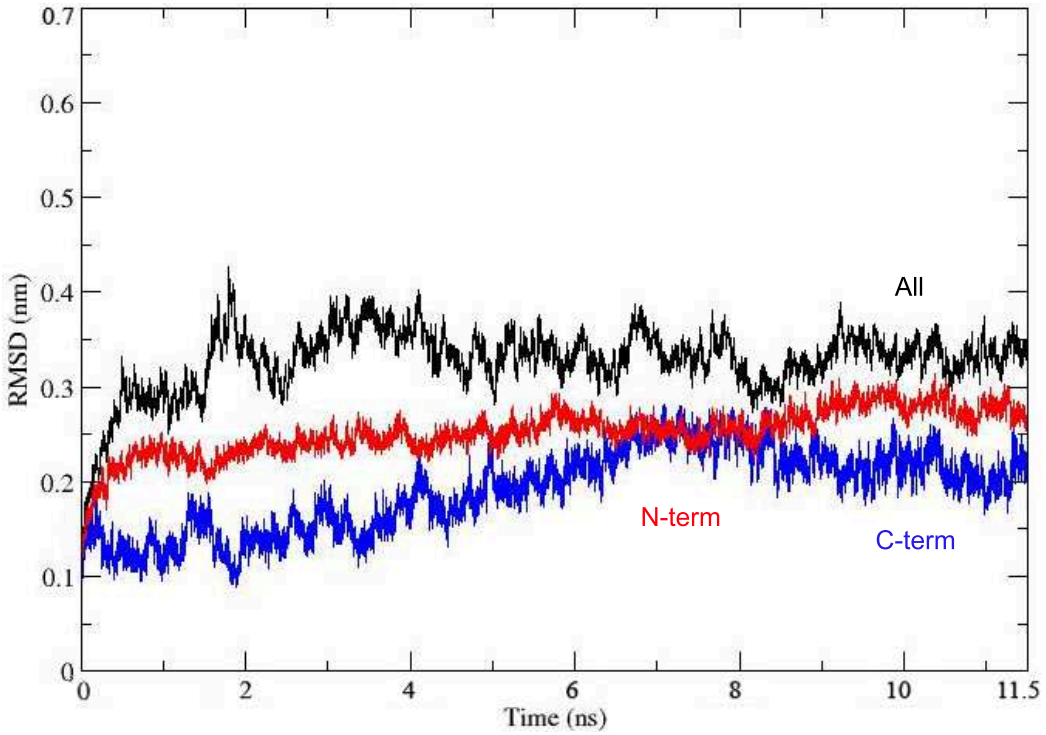


Figure 5.25: **RMSD N-terminal and C-terminal domain; CchH2-Val simulation.** Conformational drift of the combined structural components (α-helices and β-strands) of the protein in apo state CchH2, measured as C-α atom root mean square deviation (RMSD) from the starting structure. RMSDs vs. time are shown for both domains (black), the N-terminal domain (A_{core} domain) (red) and the C-terminal domain (A_{sub} domain) (blue).

5.6.2 Secondary Structure

The radius of gyration (Rg) of PheA was calculated for each simulation and plotted versus time. In each simulation the Rg of PheA decreased gradually (by between 0.05 nm^2 and 0.09 nm^2) with no significant changes observed.

In table 5.6.2 the average secondary structure content in CchH2 for each of the simulations according to DSSP classification[273], is reported. Visual plots of the evolution of the secondary structure content versus time for each simulation are included on the accompanying CD. Analysis of the secondary structure of CchH2 over the course of each of the simulations revealed a good degree of secondary structural stability in the core regions of the protein, consistent with the RMSD results. The standard deviation of the residues that form α-helices shows greater variance across the simulations; the standard deviation is greater in the CchH2-apo and CchH2-Ser simulations than in the CchH-Thr and CchH2-Val simulations, where the standard deviation is similar to that observed in the PheA1-apo and

PheA2-holo simulations. The greater variance was mainly observed in the regions not defined as part of the A domain core structure, e.g. helix H4. Overall the CchH2 structure from the holo threonine simulation exhibited the greatest degree of structural stability.

One notable region of interest within the CchH2 structure is the long loop that contains the A domain invariant lysine residue at the tip. This loop projects down from the A_{sub} domain into the A_{core} domain. In the PheA-apo simulations and PheA-holo simulations, discussed in Chapter 3, the lengths of this loop sporadically exhibited antiparallel β sheet structure on the timescale of the simulation. DSSP analysis of the CchH2-Thr simulation identified similar behaviour of this loop although to a lesser extent. This region was dominated by the presence of an antiparallel β sheet structure between 1.05 and 4.1 ns and to a lesser extent between 4.1ns and 5.65 ns, and this sheet was seen much less frequently between 6.56 and 11.5 ns. In the CchH2-apo simulation the lengths of this loop adopted an antiparallel β sheet structure very briefly at various points throughout the simulation.

5.6.3 Structural Flexibility

The time-averaged root-mean-squared fluctuations (RMSFs) for the C-α atoms of each residue in the protein provides a measure of the relative flexibility of different regions of the protein. Figure 5.26 shows the C-α atom RMSFs as a function of residue number for the CchH2-apo simulation. The greatest flexibility is exhibited by residues in the external loop regions of the protein and the N- and C-terminal linker regions of the protein. The overall pattern of fluctuations exhibited by the CchH2-apo structure C-α atoms is consistent with that of the PheA-apo simulations, however all of the fluctuations are higher.

Figure 5.27 shows the C-α atom RMSFs as a function of residue number for the CchH2-Thr simulation (red line) compared with the C-α atom RMSFs for the CchH2-apo simulation (black line). The loop region atoms exhibiting the greatest flexibility in the CchH2-apo simulation generally demonstrate reduced flexibility in the CchH2-Thr simulation; a pattern consistent with comparison of the RMSFs for the structures in the PheA-apo and PheA-holo simulations in Chapter 3. This is consistent with the CchH2 structure, which was

Secondary Structure	CchH2-apo	CchH2-Thr	CchH2-Ser	CchH2-Val
Coil	129.24 (6.59)	127.93 (6.02)	129.24 (6.59)	126.64 (7.03)
B-Sheet	118.88 (7.04)	112.86 (6.16)	118.87 (7.04)	111.28 (7.34)
B-Bridge	7.41 (2.61)	7.92 (2.93)	7.41 (2.61)	10.13 (3.09)
Bend	83.96 (6.66)	88.96 (7.69)	83.95 (6.66)	84.67 (7.67)
Turn	60.18 (8.29)	50.68 (6.60)	60.17 (8.29)	54.38 (6.93)
A-Helix	127.56 (11.46)	139.60 (6.62)	127.55 (11.46)	134.81 (5.24)
5-Helix	1.66 (2.93)	1.20 (2.50)	1.66 (2.93)	1.20 (2.27)
3-Helix	2.10 (2.47)	1.81 (2.26)	2.11 (2.47)	4.85 (3.49)

Table 5.1: **Average secondary structure contents in CchH2 apo and holo simulations.** Standard deviations are in parentheses.

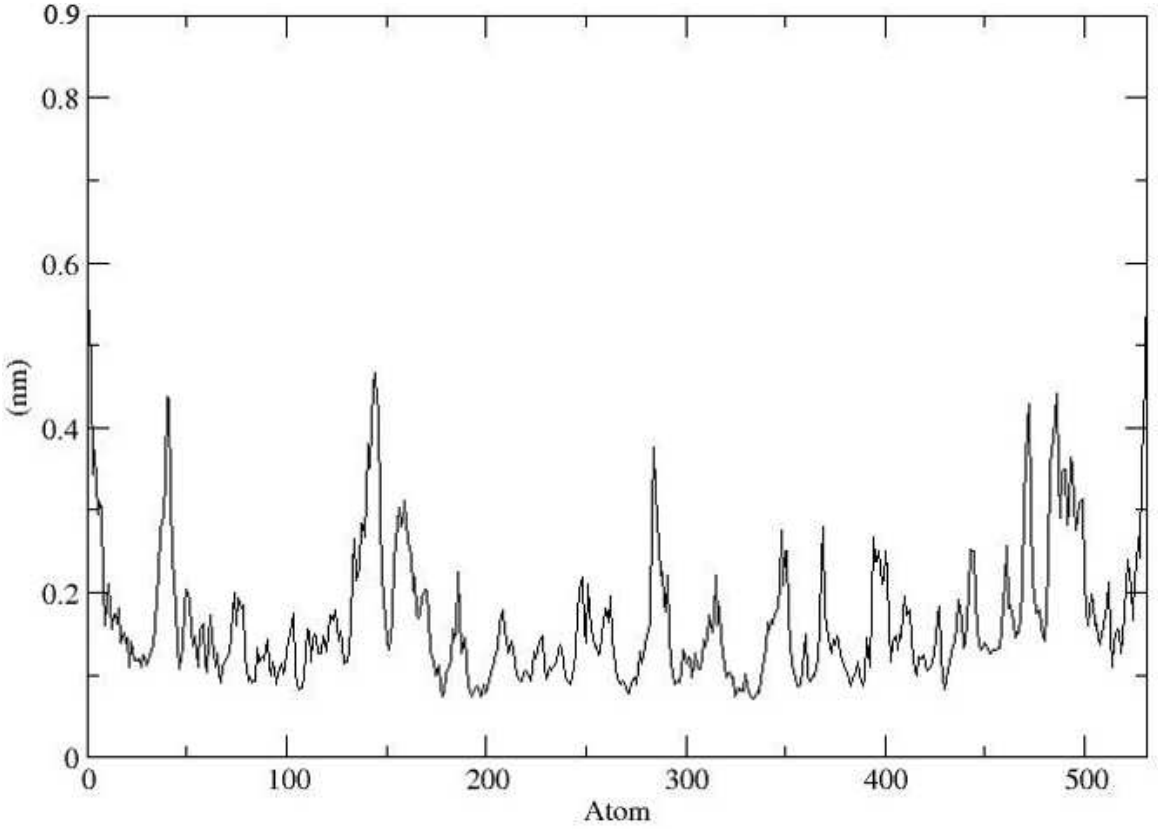


Figure 5.26: **RMSF CchH2-apo simulation.** Time-averaged C-α RMSFs as a function of residue number for the CchH2-apo state simulation.

modelled on the PheA-holo structure, being stabilised by interactions with the substrate and cofactors. The only exception is the region of CchH2 composed of residues 153 to 175 and that connects two β-strands, the latter of which contains the AMP binding motif, exhibits greater structural flexibility in the CchH2-Thr simulation than in the CchH2-apo simulation; although the shape of this region in the graph is consistent in the CchH2-thr and PheA-holo simulations. The A3 motif loop, which includes residues 182 to 187 in CchH2 and which follows directly on from the β-strand containing the AMP binding motif, displays a similar degree of flexibility in the CchH2-Thr simulation as in the CchH2-apo simulation and PheA2-holo simulation. The pattern of fluctuations demonstrated by the C-α atoms in the C-terminal domain (residues 435 to 531) is consistent with those in the equivalent region of PheA in the PheA-holo simulations; the residues which form β-strands D1 and D2 exhibit similar flexibility to those in the PheA2-holo simulation and less flexibility in the CchH2-Thr simulation than the PheA1-holo simulation.

Figure 5.28 shows the C-α atom RMSFs as a function of residue number for the CchH2-Ser

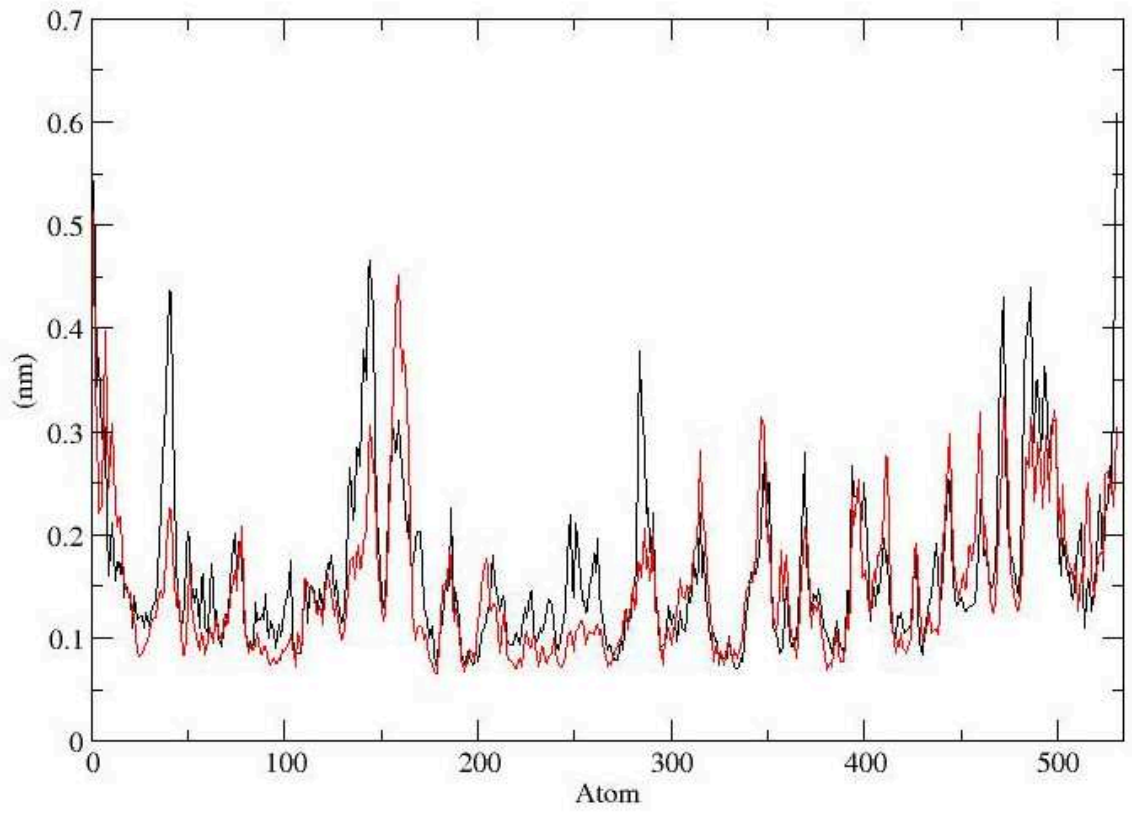


Figure 5.27: **RMSF CchH2-Thr simulation.** Time-averaged C-α RMSFs as a function of residue number for the CchH2-Thr simulation (red line) and CchH2-apo simulation (black line).

simulation (red line) compared with the C-α atom RMSFs for the CchH2-apo simulation (black line). The overall RMSF pattern for the CchH2-Ser simulation is different to that of the CchH2-Thr and PheA-holo simulations. None of the regions of structure which exhibit reduced flexibility in these holo simulations (CchH2-thr and PheA-holo) demonstrate reduced flexibility in the CchH2-Ser simulation. Interestingly the increased flexibility of residues 153 to 175 in the CchH2-Thr is not exhibited by this region in the CchH2-Ser simulation although the CchH2-Ser A3 motif loop is more flexible in the CchH2-Ser than the CchH2-Thr simulation. Three loops on the surface of the CchH2 protein formed by residues 310 to 319, 338 to 355 and 402 to 414 demonstrate a greater degree of flexibility in the CchH2-ser simulation than the CchH2-apo and CchH2-Thr simulations. The loop formed by residues 338 to 355 projects from the A_{core} domain upwards towards the A_{sub} domain and the greater degree of flexibility may be as a consequence of motion between the two domains. The pattern of fluctuations demonstrated by the C-α atoms in the A_{sub} domain (residues 435 to 531) is generally qualitatively and quantitatively consistent with that in the equivalent region of PheA in the PheA1-holo simulation; although residues 471 to 477 which form a loop demonstrate less flexibility in the CchH2-Ser simulation.

Figure 5.29 shows the C-α atom RMSFs as a function of residue number for the CchH2-Val simulation (red line) compared with the C-α atom RMSFs for the CchH2-apo simulation (black line). Some of the loop region atoms in the CchH2-Val simulation which exhibited the greatest flexibility in the CchH2-apo simulation demonstrate reduced flexibility in the CchH2-Val simulation. The increase in the flexibility of residues 153 to 175 in the CchH2-Thr simulation is seen in the CchH2-Val simulation. As is seen in the CchH2-Ser simulation the A3 motif loop (which includes residues 182 to 187) in the CchH2-Val simulation displays an increase in local flexibility when compared with the CchH2-apo simulation. In the CchH2-Val simulation the residues in α helix H5, which is located in a region of coil structure linking two β-strands, demonstrate an increase in flexibility when compared with the equivalent region in the other CchH2 simulations. Greater flexibility is displayed by external loop residues 402 to 414 in the CchH2-Val simulation than in the CchH2-apo and CchH2-Thr simulations; the flexibility in this region in the CchH2-Val simulation is comparable to that in the CchH2-Ser simulation. The pattern of fluctuations demonstrated by

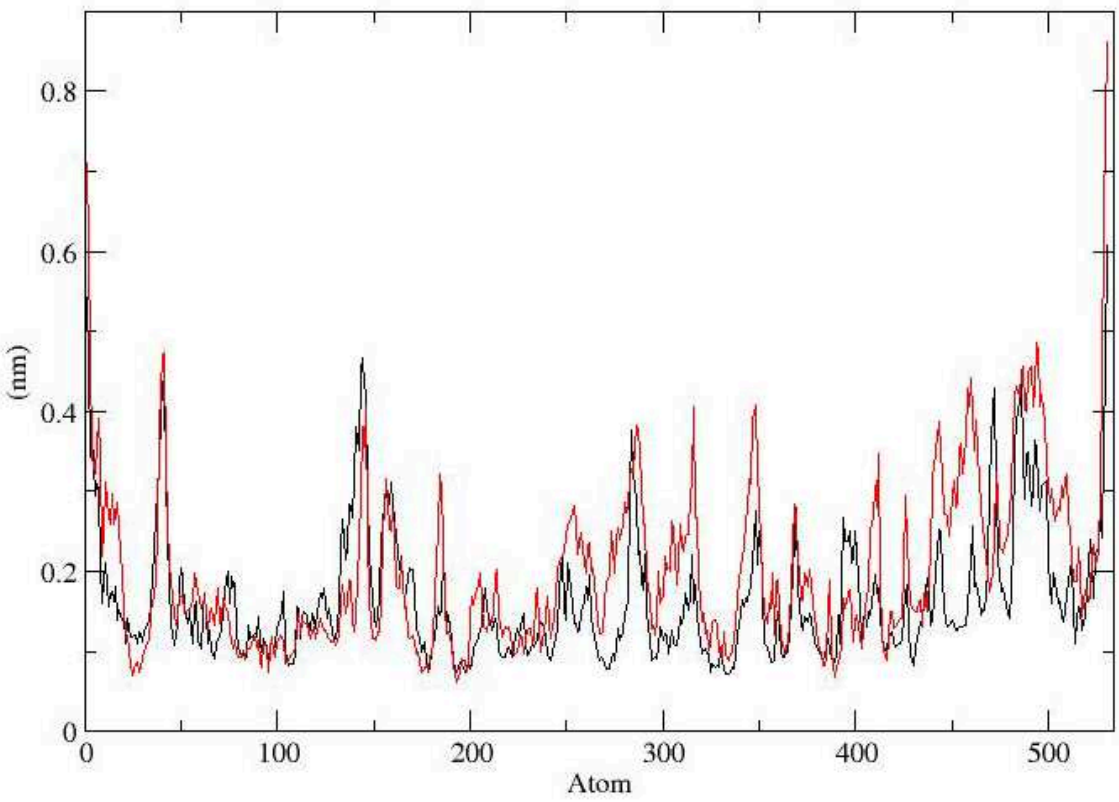


Figure 5.28: **RMSF CchH2-Ser simulation.** Time-averaged C-α RMSFs as a function of residue number for the CchH2-Ser simulation (red line) and CchH2-apo simulation (black line).

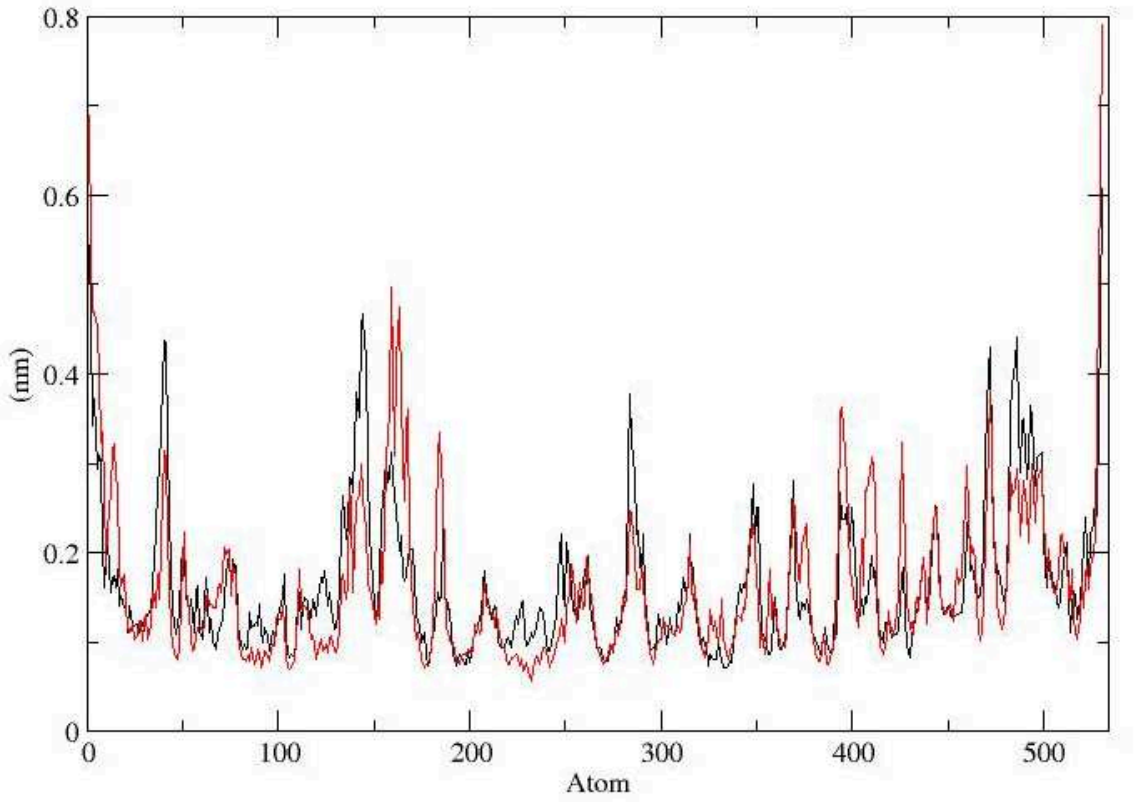


Figure 5.29: **RMSF CchH2-Val simulation.** Time-averaged C-α RMSFs as a function of residue number for the CchH2-Val simulation (red line) and CchH2-apo simulation (black line).

Index	CchH2-apo	CchH2-Thr	CchH2-Ser	CchH2-Val
1	17.07 (33.52)	9.28 (22.08)	37.6 (47.38)	17.24 (33.07)
2	5.6 (44.52)	7.78 (40.58)	10.77 (60.96)	6.46 (45.47)
3	3.8 (51.98)	3.1 (47.96)	6.22 (68.81)	3.65 (52.47)
4	1.95 (55.81)	2.66 (54.30)	4.16 (74.05)	3.11 (58.43)
5	1.87 (59.49)	1.64 (58.19)	2.49 (77.19)	2.38 (62.99)
6	1.7 (62.82)	1.26 (61.19)	1.41 (78.96)	1.63 (66.11)
7	1.15 (65.08)	1.15 (63.93)	1.15 (80.42)	1.34 (68.68)
8	1.03 (67.11)	0.89 (66.05)	1.01 (81.69)	1.01 (70.62)
9	0.85 (68.75)	0.83 (68.02)	0.86 (82.77)	0.98 (72.50)
10	0.81 (70.37)	0.72 (69.74)	0.83 (83.81)	0.9 (74.22)

Table 5.2: **PCA analysis of the CchH2 simulations**The eigenvectors (index) and eigenvalues (cumulative percentage) of the CchH2-apo, CchH2-Thr, CchH2-Ser and CchH2-Val simulations.

the C-α atoms in the C-terminal domain (residues 435 to 531) is very similar to that in the CchH2-Thr simulation.

5.6.4 Principal Modes of Motion

As for PheA, the principal modes of motion of CchH2 in each simulation were identified using PCA analysis, see table 5.2.

To analyse the nature of the collective motions of CchH2 in each system the trajectories from each principal component analysis were projected onto the respective first three eigenvectors to reveal the sampling along these vectors. The extreme projections of the trajectories along the first three eigenvectors were obtained. These structures were processed using the DynDom server. The DynDom program and visual inspection of the conformations which correspond to the extremes of the projection of the eigenvectors onto the trajectory were used to identify the nature of the any motion identified corresponding to the principal eigenvectors.

No interdomain motion was identified in CchH2 in the apo or Valine simulation. Interdomain motion was identified from the motion described by the first eigenvector in the CchH2-Thr and the first and second eigenvector in the CchH2-ser simulation.

The extremes corresponding to the motion described by the first eigenvector for the CchH2-

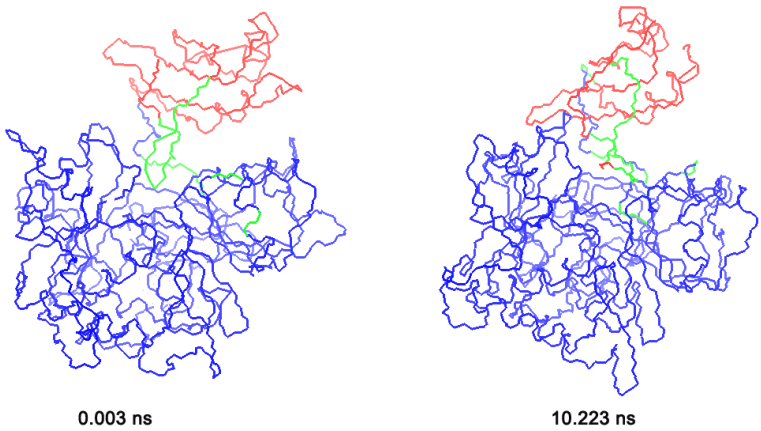


Figure 5.30: **Domain motion in CchH2-Thr.** Interdomain motion in the CchH2-Thr simulation described by the first eigenvector and identified by DynDom. Domain 1 (static) is shown in blue, domain 2 (moving) in red, the hinge regions in green.

Thr simulation are seen at 0.003 ns and 10.223 ns (see figure 5.30). DynDom analysis and overlay of the structures representing the extremes of this motion, identified the clockwise rotation (circa 35°) of subdomain E and helix H6 of the A_{sub} domain towards the A3 motif loop side of PheA. The static domain comprises the A_{core} domain and subdomain D of the A_{sub} domain and a small number of residues from the A10 motif K loop. DynDom analysis identified a number of bending residues; 301–302, 307–308, 436–436, 439–440, 447–448, 512–520 and 522–528, which include residues from the A8 and A10 motifs. Hinge residues are primarily those from the A8 motif; 436, 439 and 440.

The extremes corresponding to the motion described by the first eigenvector for the CchH2-Ser simulation are seen at 0.901 ns and 10.274 ns (see figure 5.31). This eigenvector describes the rotation of the A_{sub} domain by circa 28° towards the right side of CchH2 away from the A3 motif loop (the orientation of CchH2 used to define the right and left sides is the same as for PheA, shown in figure 3.14 of Chapter 3). As this happens the A_{sub} domain tips forward towards the binding cleft. The static domain is the A_{core} domain. A small number of residues (113–122, and 181–192 - from the A3 motif loop) from the A_{core} domain move in concert with the A_{sub} domain. DynDom analysis identified a number of bending residues; 113–114, 122–123, 181–182, 192–193 (A3 motif), 432–443 (A8 motif) and 518–519 (A10 motif). The hinge residues are primarily those from the A8 motif, 432–443, which includes the highly conserved Asp residue (437).

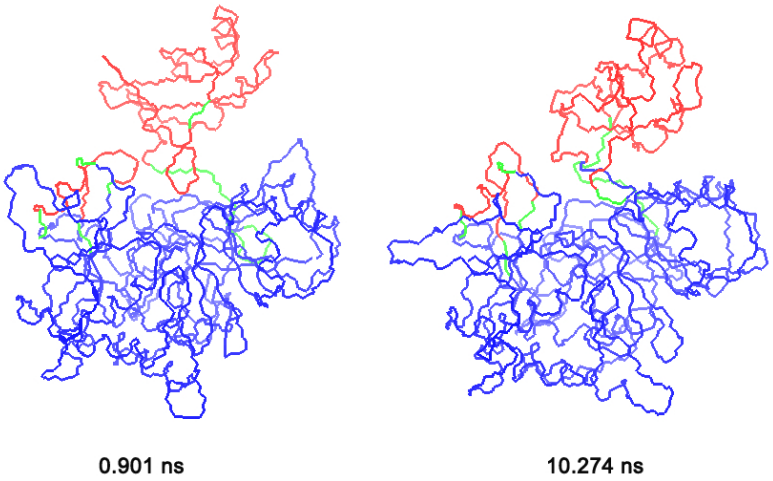


Figure 5.31: **Domain motion in CchH2-Ser eigenvector 1.** Interdomain motion in the CchH2-Ser simulation described by the first eigenvector and identified by DynDom. Domain 1 (static) is shown in blue, domain 2 (moving) in red, the hinge regions in green.

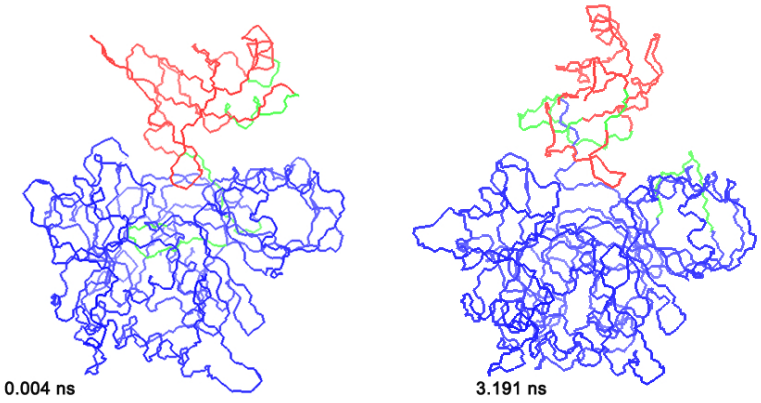


Figure 5.32: **Domain motion in CchH2-Ser eigenvector 2.** Interdomain motion in the CchH2-Ser simulation described by the second eigenvector and identified by DynDom. Domain 1 (static) is shown in blue, domain 2 (moving) in red, the hinge regions in green.

The extremes of motion from the second eigenvector are seen at 0.004 ns and 3.191 ns. This eigenvector describes the rotation of circa 38° and tilting of the A_{sub} domain towards the A3 motif loop of the A_{core} domain. A small number of residues from the A10 motif K loop are considered part of the A_{core} domain for this motion. DynDom identified a number of bending residues (342–351, 442–449, 467–475, and 509–511) including those from the A8 motif which are likely acting as a hinge for this motion. The domain motion described by eigenvector 1 of CchH2-Thr and PheA1-holo are similar, and domain motion described by eigenvector 2 of CchH2-Ser is similar to that described by eigenvector 1 of the PheA2-holo simulation.

	CchH2-apo	CchH2-Thr	CchH1-Ser	CchH2-Val
Whole simulation (SD)	683.4 (19.9)	666.1 (20.5)	671.2 (23.4)	665.4 (20.6)
1st ns (SD)	663.5 (16.0)	629.7 (17.1)	638.8 (18.5)	627.3 (18.9)
Final ns (SD)	696.3 (15.1)	668.4 (15.2)	701.0 (16.2)	673.6 (14.4)

Table 5.3: Average number of intramolecular hydrogen bonds (P-P H bonds) for the apo and holo CchH2 simulations.

5.6.5 Intramolecular Hydrogen Bonding

The average number of intramolecular (protein-protein) hydrogen bonds was obtained for CchH2 for the whole simulation, first nanosecond and last nanosecond for each simulation, see table 5.6.5. In each simulation the average number of intramolecular hydrogen bonds within CchH2 increases gradually.

5.6.6 Ligand Binding

One measure of ligand binding is an assessment of the hydrogen bonding between the ligand and protein. As for the PheA simulations, the average number of hydrogen bonds per nanosecond was calculated and will be used as a measure of the hydrogen bonding strength between particular residue groups.

Threonine Substrate

The strength of the hydrogen bonding interaction between the L-Thr substrate α-amino group and the carboxyl group of Asp 226 (equivalent to that of Asp 219 in PheA) increases during the first four nanoseconds of the simulation and then maintains an average strength 1.6 for the remainder of the CchH2-Thr simulation, see figure 5.33. A weaker yet constant hydrogen bonding interaction is formed between the L-Thr substrate α-amino group and the hydroxyl side chain of Thr 328.

The strength of the hydrogen bonding interaction between the α-carboxyl group of the Thr substrate and the amino side chain group of the invariant Lys 518 residue (PheA 501, pdb: 517) of CcH2 decreases over the time scale of the simulation, see figure 5.34, from an initial

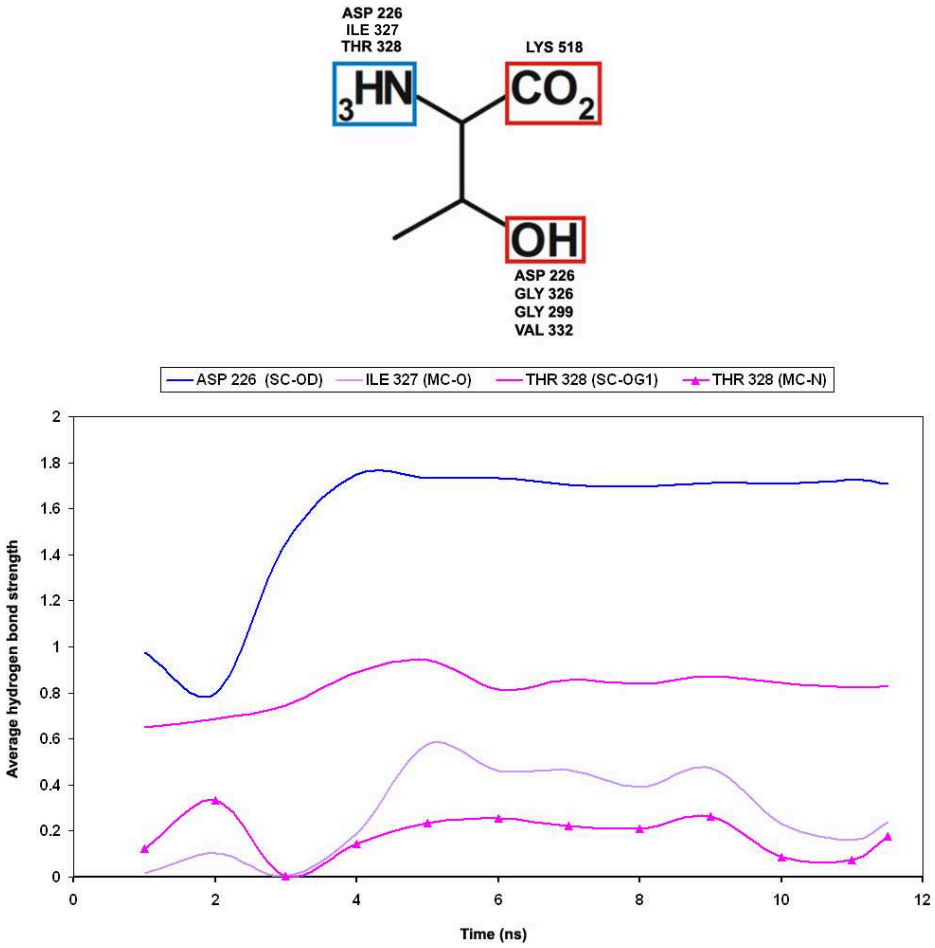


Figure 5.33: **Hydrogen bonding between L-Thr substrate and CchH2.** Top: The L-Threonine substrate annotated with the hydrogen bonding residues it interacts with. Bottom: Graph to show the hydrogen bonding interactions between the C-α amino group of the L-Threonine substrate and the CchH2 protein as a function of time. Hydrogen bonds are measured as the average strength per ns and are plotted at the ns marker.

average strength of 1.4 to 0.8 by the eleventh nanosecond.

Hydrogen bonding interactions between the hydroxyl side of the L-Thr substrate and CchH2 are few and intermittent. The hydrogen bonding interactions present between the hydroxyl side chain of Thr and carboxyl group of Asp 226 diminish in correlation with the increasing hydrogen bonding interaction strength between Asp 226 to the amino group of the L-Thr substrate carboxyl group.

Serine Substrate

The hydrogen bonding interaction between the α-amino group of the L-Ser substrate and the carboxyl side chain of Asp 226 is initially strong, see figure 5.35. The decrease in the

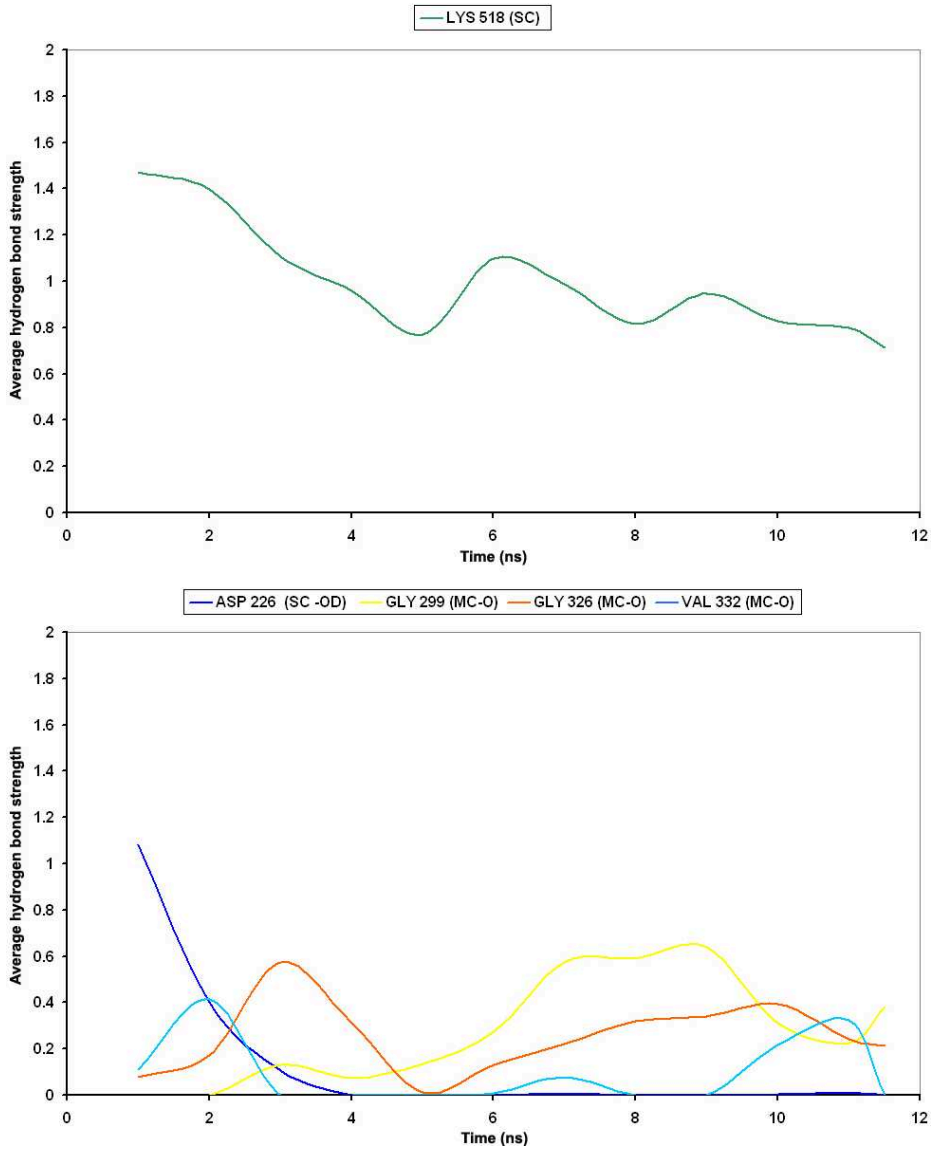


Figure 5.34: **Hydrogen bonding between L-Thr substrate sidechain and CchH2.** The hydrogen bonding interactions between the C-α carboxylate group (top graph) and the hydroxyl sidechain group (bottom graph) of the L-threonine substrate and the CchH2 protein as a function of time.

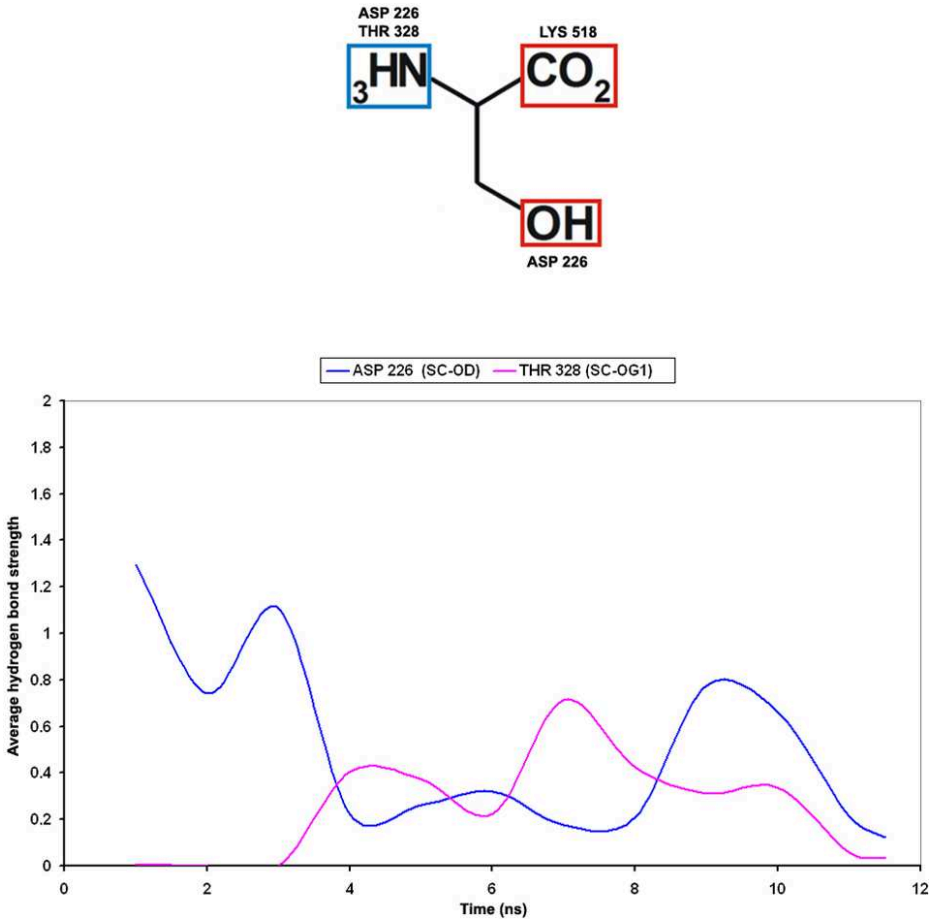


Figure 5.35: **Hydrogen bonding between L-Ser substrate and CchH2.** Top: The L-Serine substrate annotated with the hydrogen bonding residues it interacts with. Bottom: Graph to show the hydrogen bonding interactions between the C-α amino group of the L-Serine substrate and the CchH2 protein as a function of time.

strength of this interaction observed on the time scale of the simulation, correlates with the increase in the strength of the hydrogen bonding between the Asp 226 carboxyl side chain and the hydroxyl side chain of the Ser substrate, see figure 5.36. The hydrogen bonding interaction between the α-carboxyl group of the L-Ser substrate and the amino side chain group of Lys 518 is weaker (an average of 0.5) than that observed in the CchH2-Thr simulation, although a similar variation in the strength of this bonding during the simulation is apparent.

Valine Substrate

The hydrogen bonding interaction of the L-Val substrate α-amino group with the carboxyl side chain of Asp 226 is initially strong, see figure 5.37. After the third nanosecond this

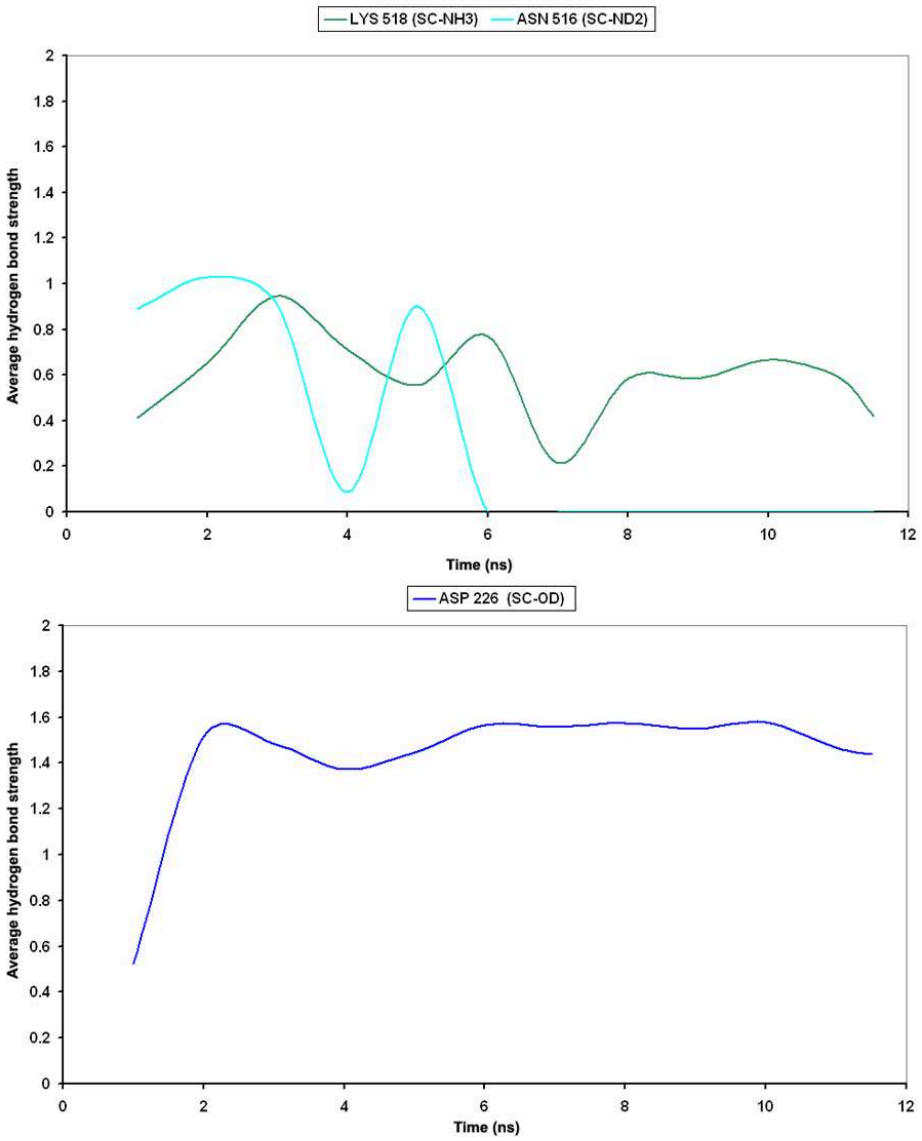


Figure 5.36: **Hydrogen bonding between L-Ser substrate sidechain and CchH2.** The hydrogen bonding interactions between the C-α carboxylate group (top graph) and the hydroxyl sidechain group (bottom graph) of the L-serine substrate and the CchH2 protein as a function of time.

interaction decreases in strength and by the end of the simulation no hydrogen bonding is present between these groups. The decrease of this hydrogen bonding interaction correlates with the decrease in strength of the hydrogen bonding between the hydroxyl side chain of Thr 328 and the Val amino group. Between the fourth and ninth nanoseconds the α-carboxyl group of the L-Val substrate forms a hydrogen bonding interaction of average strength 1.3 with the amino group of Lys 518. During the tenth nanosecond this hydrogen bonding interaction diminishes.

5.6.7 AMP Hydrogen Bonding

The hydrogen bonding interaction between the AMP ligand and CchH2 in the ChH2-Thr simulation are shown in 7.1.2 figures 7.16 and 7.17. Strong hydrogen bonding interactions are observed between the exocyclic nitrogen of the adenine moiety and the main chain groups of Met 324 and Tyr 325. As was observed in the PheA holo simulations, strong hydrogen bonding interactions are formed between the 2' and 3' hydroxyl sugars of the ribose moiety and the carboxyl side chain group of the invariant Asp residue (420). In common with what has been observed in the simulations of PheA, the binding of the phosphate group is more disordered that that of either the adenine or ribose moieties. Strong hydrogen bonding interactions are however formed between this group and the side chain hydroxyl groups of Thr 181 and Thr 328.

The hydrogen bonding interactions between the AMP ligand and CchH2 in the ChH2-Ser simulation are shown in 7.1.2 figures 7.18 and 7.19. Weaker hydrogen bonding interactions are observed between the exocyclic nitrogen of the adenine moiety and the main chain groups of Met 324 and Tyr 325, than are seen in the CchH2-Thr simulation. The hydrogen bonding interactions formed between the 2' and 3' hydroxyl sugars of the ribose moiety and the carboxyl side chain group of the invariant Asp residue (420) are also weaker and more variable in the CchH2-Ser simulation than the CchH2-Thr simulation. The binding of the phosphate group is similarly disordered in the CchH2-Ser simulation as in the CchH2-Thr simulation. The strong hydrogen bonding interactions between the phosphate group and the side chain hydroxyl groups of Thr 181 and Thr 328 as seen in the CchH2-Thr simulation

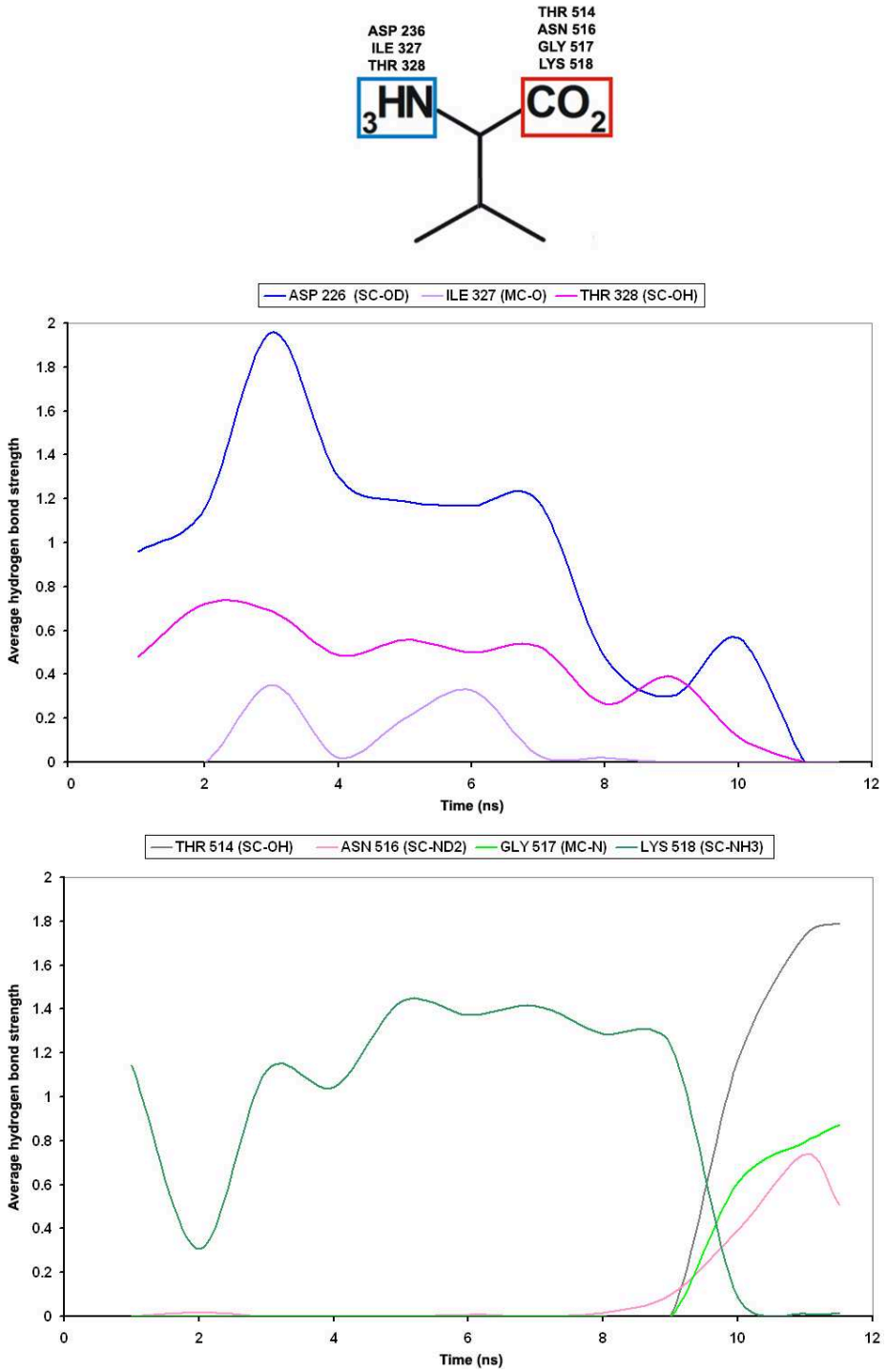


Figure 5.37: **Hydrogen bonding between L-Val substrate and CchH2.** Top: The L-Valine substrate annotated with the hydrogen bonding residues it interacts with. Middle: Graph to show the hydrogen bonding interactions between the C-α amino group of the L-Valine substrate and the CchH2 protein as a function of time. Bottom: Graph to show the hydrogen bonding interactions between the C-α carboxylate group of the L-Valine substrate and the CchH2 protein as a function of time.

are weaker and fluctuate more in the CchH2-Ser simulation.

The hydrogen bonding interactions between the AMP ligand and CchH2 in the ChH2-Val simulation are shown in 7.1.2 figures 7.20 and 7.21. Hydrogen bonding interactions of a similar strength and duration to that observed in the CchH2-Thr simulation, are seen between the exocyclic nitrogen of the adenine moiety and the main chain groups of Met 324 and Tyr 325 in the CchH2-Val simulation. In addition the hydrogen bonding interactions formed between the 2' and 3' hydroxyl sugars of the ribose moiety and the carboxyl side chain group of the invariant Asp residue (420) in the CchH2-Val simulation are of comparable magnitude and strength to those observed in the CchH2-Thr simulation. The hydrogen bonding interactions formed between the phosphate group of AMP and CchH2 are less disordered in the CchH2-Val simulation than in any of the other CchH2 holo simulations. A single hydrogen bond formed is formed between the O1P phosphate oxygen and the main chain amino group of Thr 328 and the O3P phosphate oxygen atoms and the side chain hydroxyl of Thr 328.

5.6.8 Magnesium ion coordination

The magnesium ion in each of the CchH2-holo systems is coordinated to six oxygen atoms. In all the CchH2-holo simulations the ligands to the Mg^{2+} are the carboxylate of Glu 329 (atoms OE1 and OE2), two oxygens of the AMP phosphate (O1P and O2P) and two water molecules (OW1 and OW2). The mean and standard deviation for the bond lengths (Mg-O) and angles (O-Mg-O) calculated, using data collected every ps, over the first and last ns, and the entire simulation are listed in the tables in appendix 7.1.2 figures 7.22, 7.23 and 7.24 for the CchH2-thr, CchH2-ser and CchH2-val simulations respectively. Each mean bond length (Table 1) varies only slightly on the simulation timescales (the maximum standard deviation value is 0.007) and the bond lengths between equivalent atoms are comparable when comparing the geometries across the simulations.

The bond angles (Table 2) exhibit a greater degree of variation in each system as reflected by the standard deviation value for each angle calculated over the entire simulation. An

overall trend emerges when comparing the angles in the Mg-ligand octahedral complexes in the three systems. Unsurprisingly the angles between atoms which are joined and the Mg ion, OE1-Mg-OE2 and O2P-Mg-O1P, exhibit the least variation on the simulation timescale having standard deviation values of 1.64 and 1.75 respectively in the CchH2-Thr system, 1.62 and 1.77 respectively in the CchH2-Ser system and 1.64 and 1.76 respectively in the CchH2-Val system. The angles that exhibit the greatest degree of fluctuation around the mean value over the entire simulation are the same for all systems; OE1-Mg-O2P and O1P-Mg-OE2. Both of these angles are between an atom in the AMP phosphate group, the Mg ion and an atom in the Glu 329 residue of CchH2. With a few minor exceptions the standard deviation of each angle decreases when comparing the values of the first and last ns.

The RMSD of the Glu 329 CchH2 residue has been calculated over the timescale of the simulation for each system after least squares fitting to both self and the AMP molecule. The RMSD of Glu 329, averaged over the whole simulation, after least squares fitting to self is 0.026, 0.045, and 0.031 nm, and after least squares fitting to the AMP molecule 0.088, 0.268 and 0.094 nm for the CchH2-Thr, CchH2-Ser and CchH2-Val systems respectively. The Glu 329 residue, with respect to both itself and the AMP molecule in the starting structure, fluctuates more in the CchH2-Ser system than in the CchH2-Thr and CchH2-Val system simulations.

Overall the octahedral geometry of the Mg-ligand complex is well maintained in each system simulation suggesting the force field is well parameterised for the inclusion of the Mg ion. The positioning of the Mg ion and the bond lengths it forms and maintains with the AMP and Glu 329 oxygen ligands in each of the CchH2-holo systems provides evidence of the potential incorrect positioning of the Mg ion in the PheA crystal structure file.

5.6.9 Conclusions

A homology model of CchH2 was constructed that showed good stability in the core regions of structure on the timescale of the simulations despite the protein having only $\sim$30% sequence identity with the template structure. Some α-helices that do not form part of the core

A domain structure showed greater structural variation throughout the simulations. These α-helices are mainly located in the larger A_{core} domain which showed greater structural drift in the CchH2 simulations as compared to the PheA simulations.

While no interdomain motion was identified in the CchH2-apo or CchH2-Val simulation, interdomain motion was identified in the CchH2-Thr and CchH2-Ser simulations. The first eigenvector from the CchH2-Thr simulation describes rotation of subdomain E and helix H6 of the A_{sub} domain relative to the A_{core} domain and subdomain D of the A_{sub} domain. In this motion subdomain E and helix H6 of the A_{sub} domain move towards the A3 motif loop side of PheA. The division of the domains between which the motion occurs in the CchH2-Thr simulation and the direction of the motion is consistent with that described by the first eigenvector of the PheA1-holo simulation. This motion widens the opening between the domains on the right side of the A domain enlarging an opening through which the PPant arm of the PCP domain could pass to carry out the second half reaction.

The motion described by eigenvector 1 from the CchH2-Ser simulation, rotation of the A_{sub} domain towards the right side of the A_{core} domain away from the A3 motif loop and forward towards the binding cleft, is different to that observed in the PheA simulations. However the motion described by the second eigenvector is similar to that described by eigenvector 1 of the PheA2-holo simulation.

The substrate in the cognate substrate simulation, CchH2-Thr, is more strongly bound to CchH2 than either the L-Ser or L-Val substrates; as assessed by the average hydrogen bonding. The stronger interaction of the L-Thr ligand is accompanied by a stronger interaction of the AMP ligand. In contrast, while the binding of AMP in the CchH2-Ser is more disordered and weaker, binding of AMP to CchH2 in the CchH2-Val simulation is comparable with that seen in CchH2-Thr. Ranking of the preference of CchH2 for the substrate ligand based on the hydrogen bonding strength and patterns observed between the substrate and CchH2 in each simulation places the preferred CchH2 ligand as L-Thr, followed by L-Val and L-Ser. Interestingly the L-Val substrate forms stronger hydrogen bonding interactions with CchH2, even though L-Ser is of a more similar chemical nature to L-Thr than L-Val.

It is interesting to note that interdomain rotation is only observed in some of the holo simulations and that motion is observed in the simulation with L-Ser, which ranked the lowest in an assessment based on an assessment of hydrogen bonding interactions, and not in the CchH2-Val simulation. As observed in the PheA non cognate L-Tyr simulation the presence of the hydrogen bonding interactions between the substrate and the key binding pocket residues (Asp and Lys) seem to be required for the domain rotation towards the A3 motif loop which increases the widening between the domains on the right side. This motion in the CchH2-Ser simulation is described by the second eigenvector the extreme motion of which is observed at 3.191 ns. Until this time the L-Ser substrate forms interactions with the Asp and Lys key binding pocket residues. During the fourth nanosecond, each of these interactions decreases. The extreme motion of the first eigenvector, which describes a different motion than observed in PheA, occurs later in the simulation at 10.274 ns.

These results, that demonstrate that the homology model shows good stability in the core regions of structure on the timescale of the simulation and that the L-Thr specific CchH2 A domain binds the L-Thr substrate well and better than non-cognate substrates, suggest that homology modelling of the A domains may be a useful technique for further study of the dynamics and substrate interactions of the A domains.

Chapter 6

Other Studies

6.1 Introduction

This chapter is split into two sections. In the first section an initial MD study of the PheA A domain with point mutations to attempt to confer preference of the domain from L-Phe to L-Asp are presented. In the second section the design of the set up of metadynamics calculations to characterise the free energy of the domain rotation observed in PheA is briefly described.

6.2 Point Mutation Simulations

To attempt to increase the binding affinity of PheA for the Asp substrate a preliminary study of a set of mutations in PheA was performed. Two separate point mutations to the active site were introduced to PheA and tested by performing simulations, of length 7.5 ns, on each single point mutations protein with the Phe and Asp ligands. The location and nature of these mutations was guided by the sequence analysis work of Stachelhaus *et. al*[81] and Challis and Ravel[82]. The proposed specificity conferring code for the L-Phe and L-Asp substrate is shown in table 6.1. The mutations tested were; Thr 262 mutated to Lys, and Ala 306 to His.

Substrate	235	236	239	278	299	301	322	330	331	517
Phe	D	A	W	T	I	A	A	I	C	K
Asp	D	L	W	K	V/I	G	H/A	I/V	G	K

Table 6.1: Comparison of the L-Phe and L-Asp Adenylation domain binding pocket specificity conferring code.

These simulations are referred to as PheA-Phe-Lys (Phe ligand Lys mutation), PheA-Asp-Lys, PheA-Phe-His, and PheA-Asp-His. The analysis of these simulations focuses primarily on the interdomain motion and hydrogen bonding of PheA to the ligands.

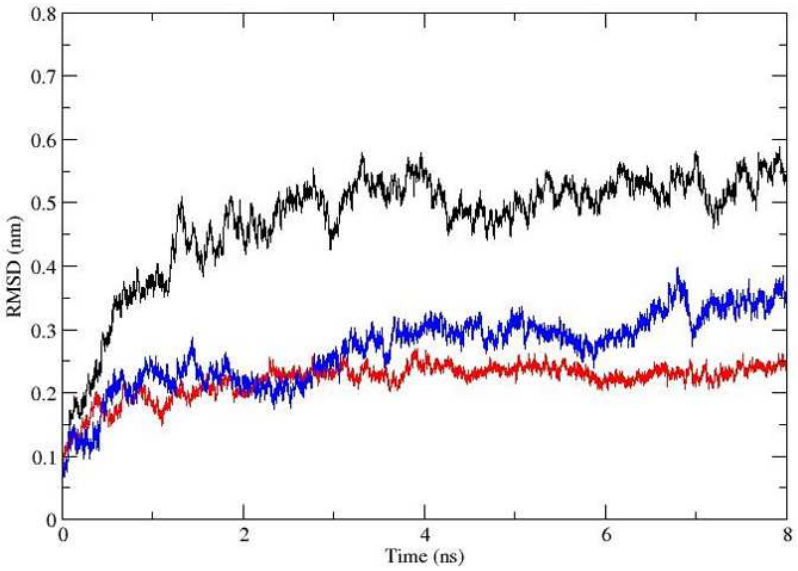


Figure 6.1: **RMSD PheA-Phe-Lys simulation.** The conformational drift of holo state PheA (PheA2-holo), measured as Cα atom root mean square deviation (RMSD) from the starting structure. RMSDs vs. time are shown for the entire protein (black), the A_{core} domain (red) and A_{sub} domain (blue).

6.2.1 Simulation Set Up

The binding pocket residues and substrate were mutated from their original amino acid to the required amino acid using Swiss-Pdb Viewer[302].

The PheA mutation Phe and Asp systems were subjected to up to 100 steps of steepest descent minimisation with all heavy atoms tethered to their original position. After the addition of solvent (water and counterions), up to 100 steps of steepest descents minimisation was performed with all heavy atoms tethered to their original position. Following this, 100 steps of conjugant gradients minimisation was performed with only the heavy atoms of the substrate (either L-Phe or L-Asp) and the ten residues lining the binding pocket tethered using a harmonic potential with a force constant of 500 kJ mol^{-1} nm^{-2} . This was followed by up to a further 50 steps of steepest descents and up to 50 steps of conjugant gradients unrestrained minimization. The NVT and NPT simulations were carried out using the protocols outlined in Chapter 4.

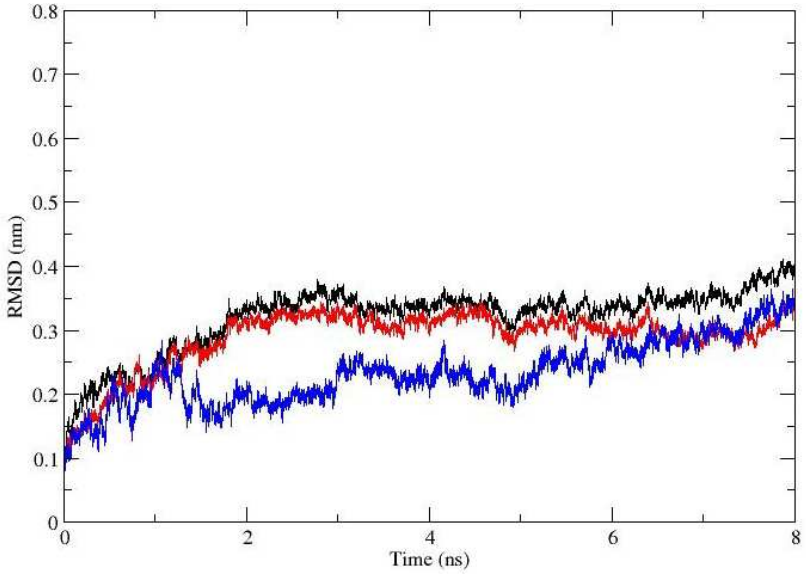


Figure 6.2: **RMSD PheA-Asp-Lys simulation.** The conformational drift of holo state PheA (PheA2-holo), measured as Cα atom root mean square deviation (RMSD) from the starting structure. RMSDs vs. time are shown for the entire protein (black), the A_{core} domain (red) and A_{sub} domain (blue).

6.2.2 Structural Drift

As in the PheA holo, and PheA-Tyr simulations the RMSD of the entire protein (black), A_{core} domain (red) and A_{sub} domain (blue) from the PheA-Phe-Lys simulation, figure 6.1 reveals the A_{core} domain to be the most structurally stable, followed by the A_{sub} domain. The greatest structural drift is observed in the structure of the entire protein.

In contrast to the observations from the PheA-Phe-Lys simulation, the RMSD of the entire protein (black), A_{core} domain (red) and A_{sub} domain (blue) from the PheA-Asp-Lys simulation, figure 6.2 reveals the A_{sub} domain to be the most structurally stable, followed by the A_{core} domain. Comparable structural drift is observed in the entire protein and the A_{core} domain.

In the initial stages of the PheA-Phe-His simulation, see figure 6.4 the A_{sub} domain exhibits the greatest structural stability. From 2 nanoseconds onwards greater structural drift is observed in the A_{core} domain

As in the PheA holo, and PheA-Tyr simulations the RMSD of the A_{core} domain (red) in the PheA-Asp-His simulation, figure 6.4, reveals this region to be the most structurally stable.

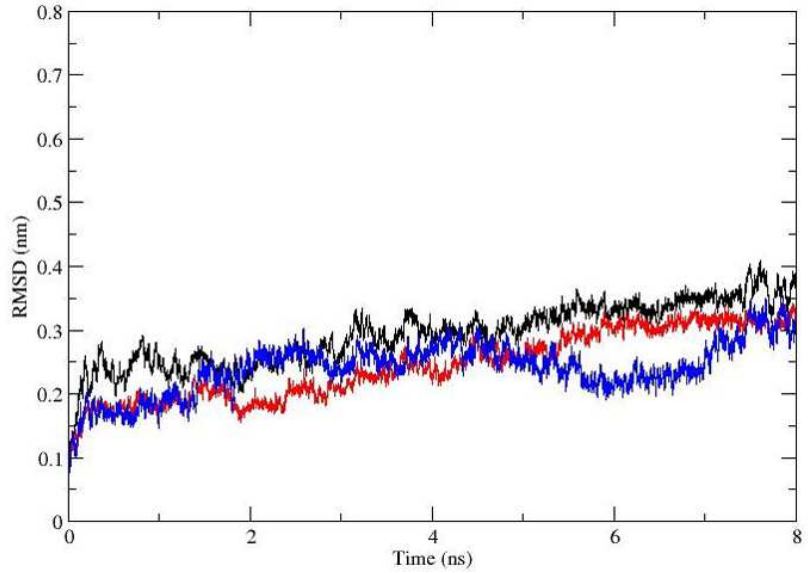


Figure 6.3: **RMSD PheA-Lys-Phe simulation.** The conformational drift of holo state PheA (PheA2-holo), measured as Cα atom root mean square deviation (RMSD) from the starting structure. RMSDs vs. time are shown for the entire protein (black), the A_{core} domain (red) and A_{sub} domain (blue).

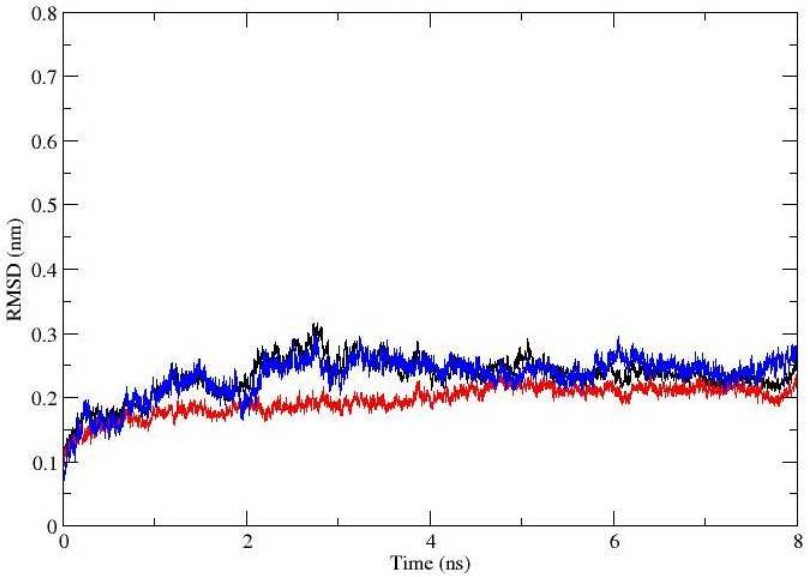


Figure 6.4: **RMSD PheA-Asp-His simulation.** The conformational drift of holo state PheA (PheA2-holo), measured as Cα atom root mean square deviation (RMSD) from the starting structure. RMSDs vs. time are shown for the entire protein (black), the A_{core} domain (red) and A_{sub} domain (blue).

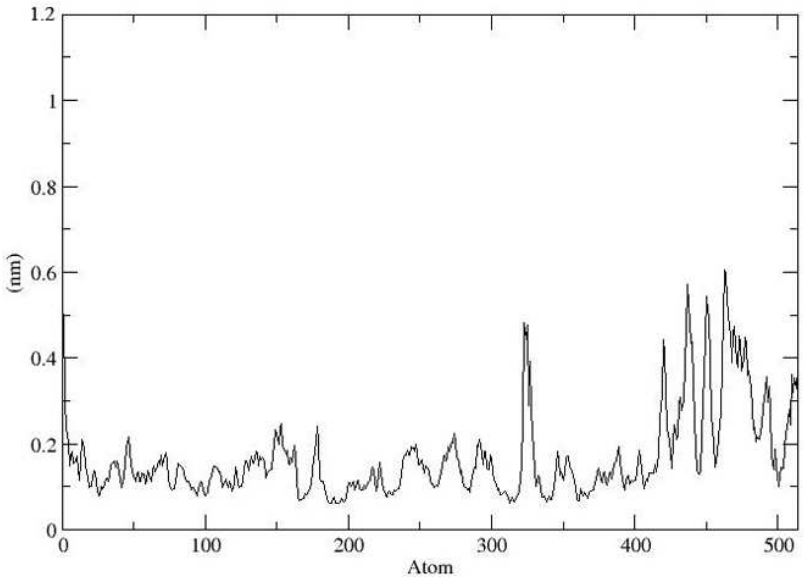


Figure 6.5: **RMSFs of the PheA-Phe-Lys simulation.** The time-averaged Cα RMSFs as a function of residue number for the PheA-Phe-Lys simulation.

The RMSDs of the A_{sub} domain (blue) and entire protein are of comparable magnitude and evolution.

6.2.3 Residue by Residue Fluctuations

The time-averaged Cα RMSFs as a function of residue number for the PheA-Phe-Lys simulation are shown in figure 6.5. This analysis reveals greater fluctuations of the A_{sub} domain as compared to the A_{core} domain. In contrast fewer fluctuations are observed in the A_{sub} domain of PheA-Asp-Lys, figure 6.6, than are observed in the PheA-Phe-Lys simulation. The fluctuations observed in this region for PheA-Asp-Lys are still marginally higher than those observed in the A_{core} domain.

The time-averaged Cα RMSFs as a function of residue number for the PheA-Phe-His simulation are shown in figure 6.7. This analysis reveals comparable fluctuations of the A_{sub} domain as compared to the A_{core} domain, and this trend is similarly observed in the time-averaged Cα RMSFs as a function of residue number for the PheA-Asp-His simulation, shown in figure 6.8.

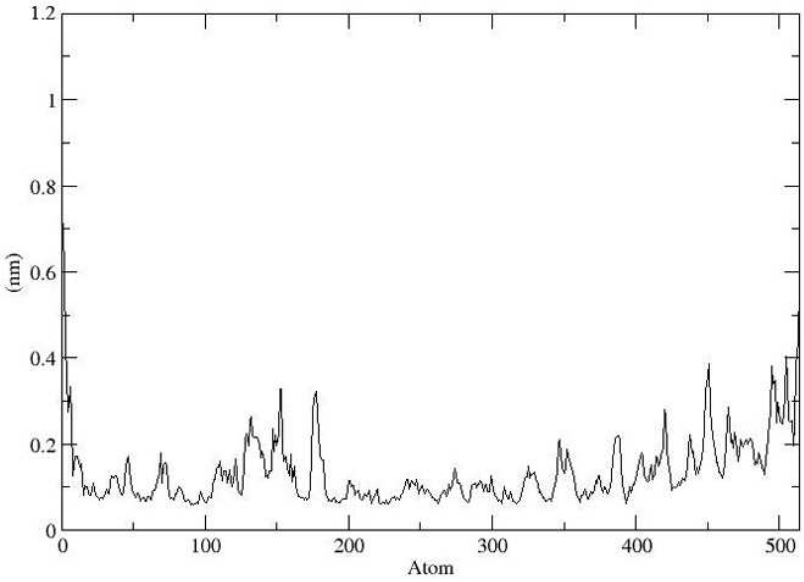


Figure 6.6: **RMSFs of the PheA-Phe-Lys simulation.** The time-averaged Cα RMSFs as a function of residue number for the PheA-Phe-Lys simulation.

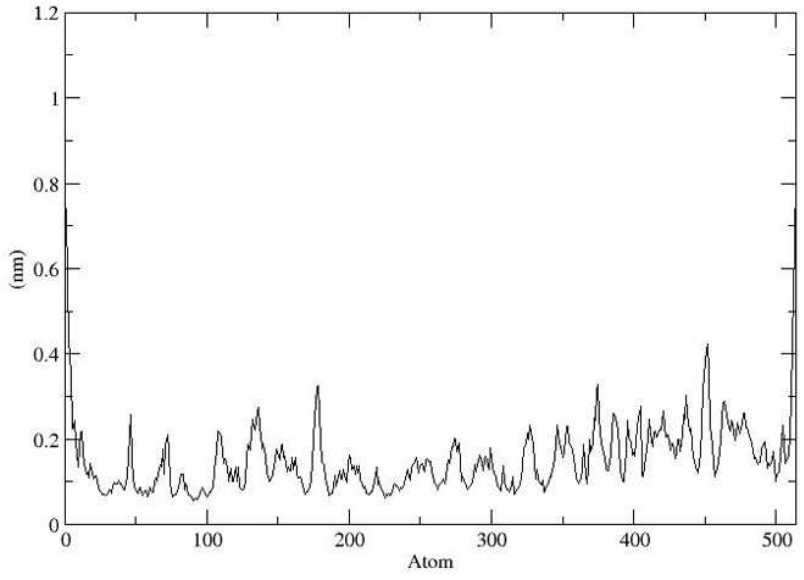


Figure 6.7: **RMSFs of the PheA-Phe-His simulation.** The time-averaged Cα RMSFs as a function of residue number for the PheA-Phe-Lys simulation.

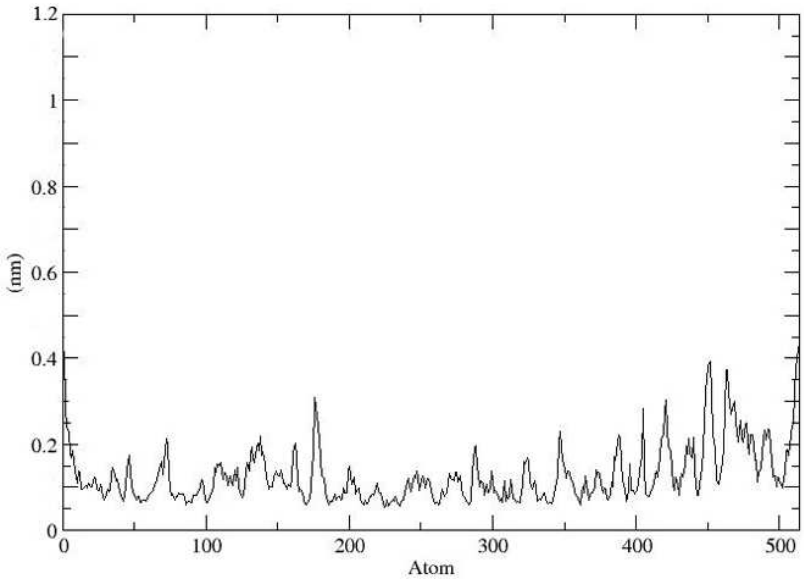


Figure 6.8: **RMSFs of the PheA-Phe-His simulation.** The time-averaged Cα RMSFs as a function of residue number for the PheA-Phe-Lys simulation.

6.2.4 Principal Modes of Motion and DynDom Analysis

The principal modes of motion of PheA were established from each mutation simulation using PCA analysis; figure 6.9 describes the size of each of the ten first eigenvectors (index). The DynDom program and visual inspection of the conformations which correspond to the extremes of the projection of the eigenvectors onto the trajectory were used to establish the nature of motion corresponding to the principal eigenvectors.

Table 6.2 summarizes the interdomain motion identified in the first two eigenvectors of the PheA-Phe-Lys simulation. Eigenvector 1 describes clockwise twisting and tilting of the A_{sub} domain towards the right side of PheA, as illustrated in Ev1 of figure 6.10. Eigenvector 2 describes anticlockwise twisting and tilting of the A_{sub} domain towards the right side of PheA, as illustrated in Ev2 of figure 6.10.

The motion identified by domain the first eigenvector of the PheA-Asp-Lys simulation describes the lifting of the A10 motif K loop towards the front of the binding pocket and towards the right side of PheA, away from the A3 motif loop, figure 6.11. No discernible interdomain motion was determined for the second and third eigenvectors of PheA-Asp-Lys. The extremes of the motion of the first three eigenvectors of the PheA-Asp-Lys simulation are observed at; 0.006 and 7.6089 ns, 0.005 and 3.001, and 2.120 and 7.498 ns respectively.

	PheA-Phe-Lys		PheA-Asp-Lys		PheA-Phe-His		PheA-Asp-His	
index	Ev (nm2)	Cumulative %	Ev (nm2)	Cumulative %	Ev (nm2)	Cumulative %	Ev (nm2)	Cumulative %
1	30.08	53.05	12.83	36.20	17.89	40.30	9.68	32.18
2	8.80	68.57	4.03	47.56	5.80	53.38	4.41	46.86
3	2.64	73.22	2.69	55.13	2.72	59.51	2.19	54.13
4	1.74	76.29	2.22	61.41	2.07	64.17	1.37	58.67
5	1.27	78.54	1.59	65.90	1.57	67.69	0.98	61.95
6	0.89	80.11	0.87	68.34	1.27	70.56	0.82	64.67
7	0.62	81.20	0.68	70.27	0.92	72.63	0.63	66.76
8	0.51	82.10	0.60	71.97	0.86	74.56	0.56	68.61
9	0.47	82.94	0.46	73.27	0.61	75.94	0.46	70.15
10	0.44	83.71	0.41	74.41	0.52	77.11	0.41	71.50

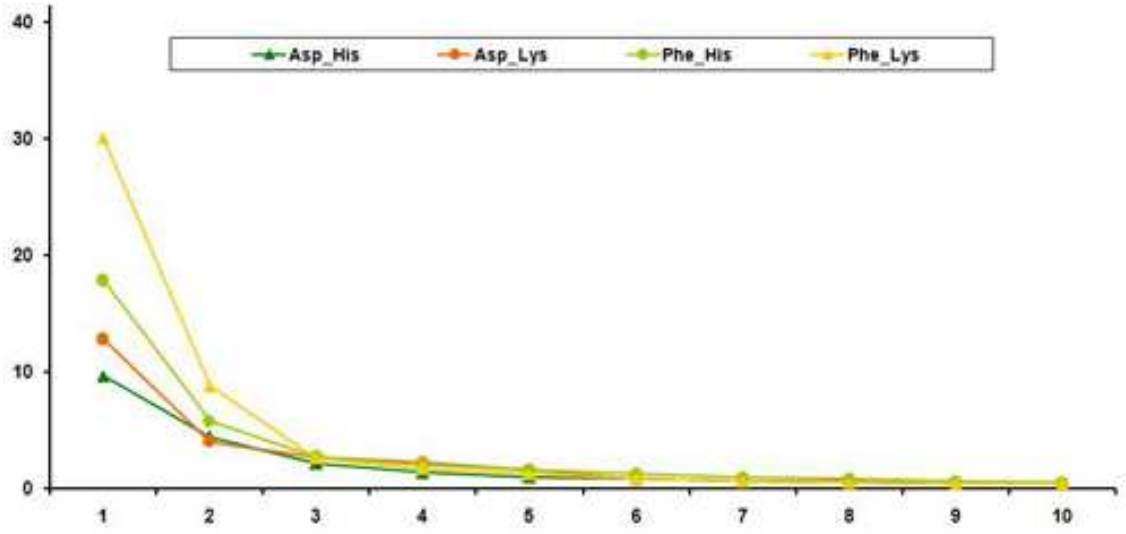


Figure 6.9: **PCA analysis of the PheA-Phe-Lys, PheA-Asp-Lys, PheA-Phe-His and PheA-Asp-His simulations.** The eigenvectors (index) and eigenvalues of the PheA-Phe-Lys (yellow), PheA-Asp-Lys (orange), PheA-Phe-His (lime) and PheA-Asp-His (green) simulations.

Ev	Extremes (ns)	Domain 1	Domain 2
1	0.493–6.161	3–176,181–217,224–415	177–180, 218–223, 416–512
2	0.003–2.595	6–173,184–411	174–183, 412–509

Table 6.2: Summary of the domain motion identified by DynDom from the first two eigenvectors of the PheA-Phe-Lys simulation.

Ev	Rotation (°)	Translation (Å)
1	40.5	1.4
2	54.3	1.4

Table 6.3: Summary of the magnitude of the domain motion identified by DynDom from the first two eigenvectors of the PheA-Phe-Lys simulation.

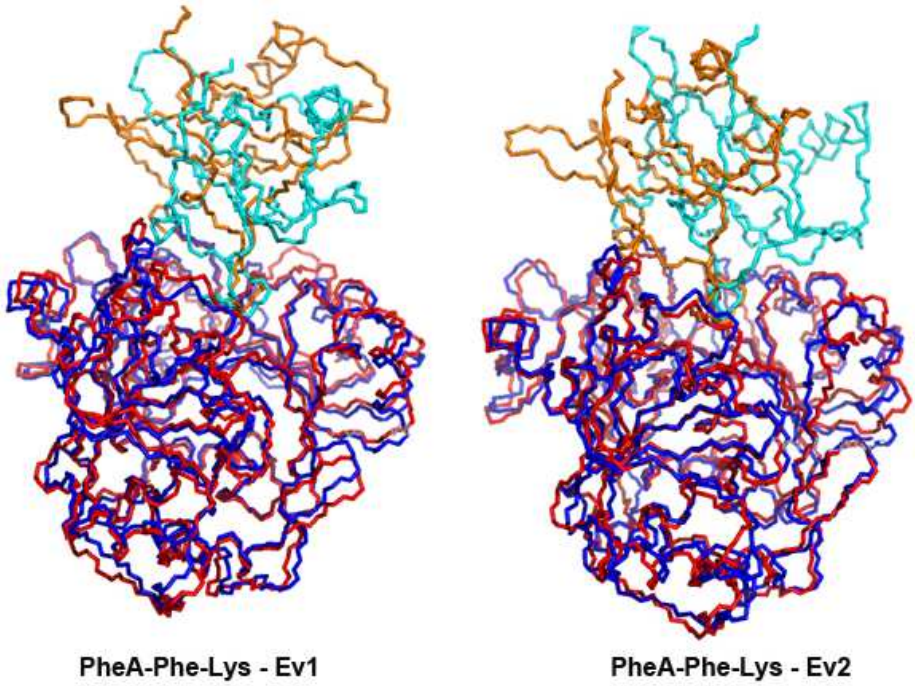


Figure 6.10: **Visualisation of the motion of the first two eigenvectors of the PheA-Phe-Lys simulation.** The initial structure (red/orange) is shown overlayed with the final structure (blue/cyan).

No discernible interdomain motion was determined for the principal eigenvectors of PheA-Phe-His simulation using DynDom. DynDom did however identify motion in the first eigenvector of the PheA-Asp-His simulation, see figure 6.12. This motion was evident at its extreme between 1.002 and 7.109 ns. Essentially this motion describes the upright anti-clockwise twisting of the A_{sub} domain. DynDom identified the static domain (domain 1) in this motion as being comprised of residues 3–285, 290–416, 419–422, 434–439, 447–449, and 479–481, and domain 2 of residues 286–289, 417–418, 423–433, 440–446, 450–478, and 482–512. Residues 285–286, 289–290, 415–423, 433–434, 436–440, 446–450, and 478–482 were identified as bending hinge residues.

6.2.5 Hydrogen Bonding of Substrate with PheA

As this is a preliminary study the analysis performed on the trajectories of these simulations are not as extensive as that performed for the apo and holo simulations and noncognate ligand simulations. One measure of ligand binding is an assessment of the hydrogen bonding between the ligand and protein. The average number of hydrogen bonds per nanosecond

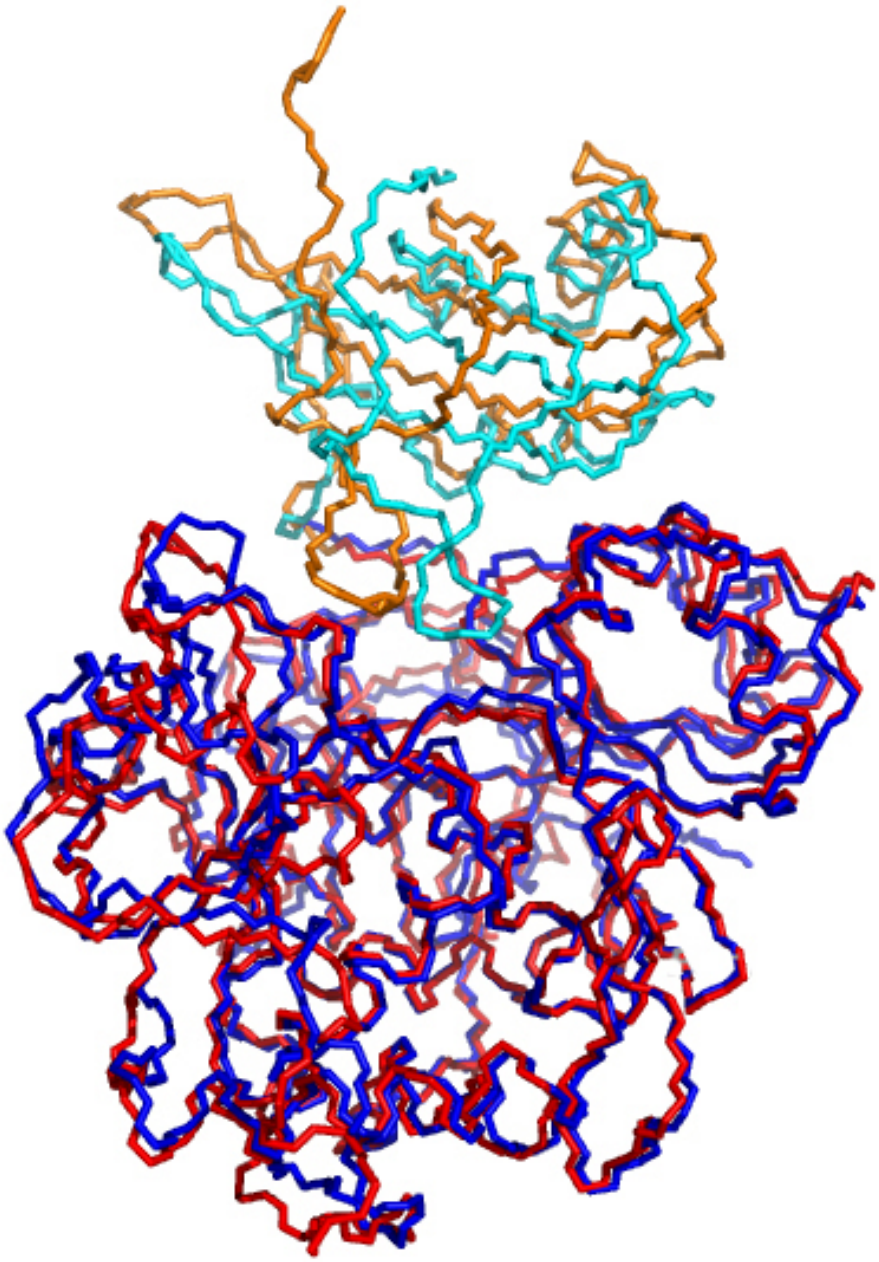


Figure 6.11: **Visualisation of the motion of the first eigenvector of the PheA-Asp-Lys simulation.** The movement from the initial structure (red/orange) is shown overlayed with the final structure (blue/cyan).

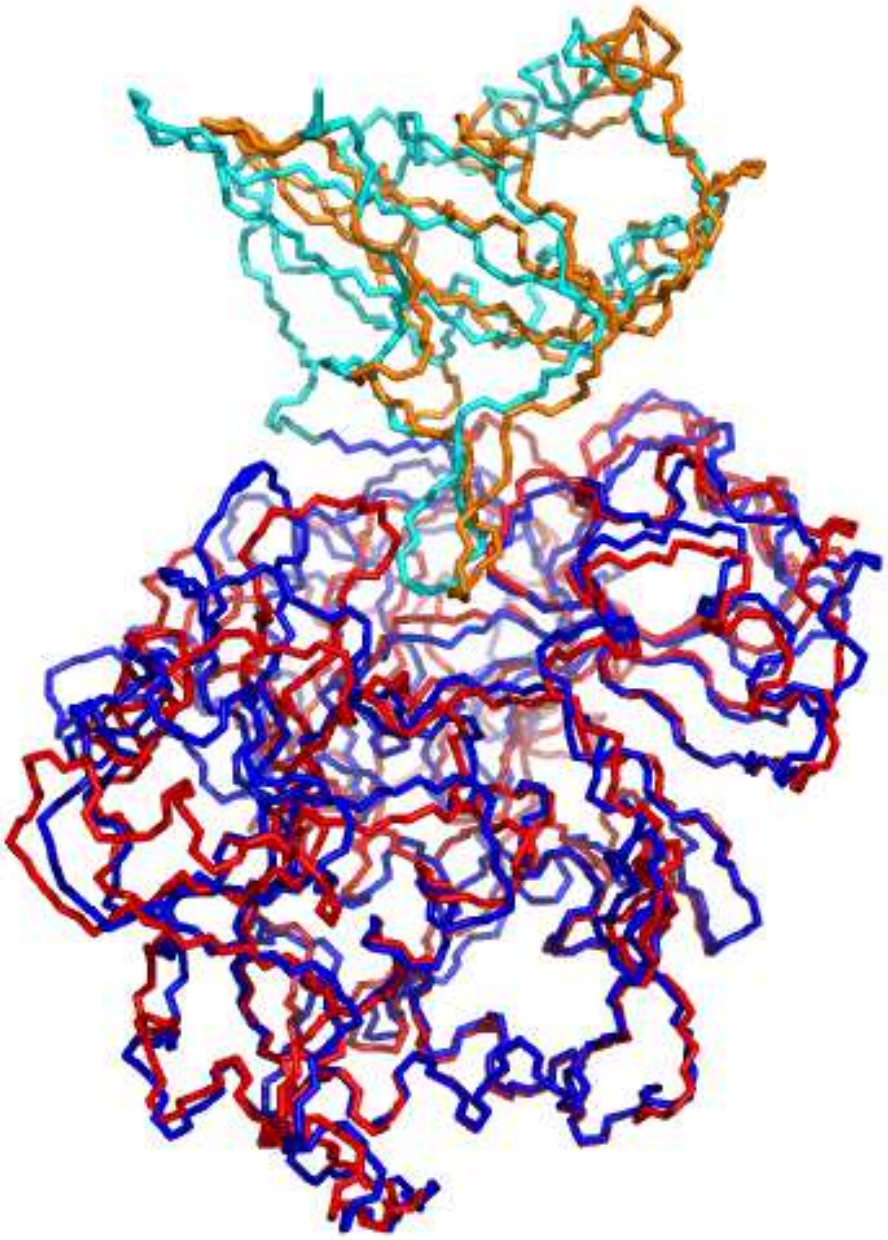


Figure 6.12: **Visualisation of the motion of the first eigenvector of the PheA-Asp-His simulation.** The movement from the initial structure (red/orange) is shown overlayed with the final structure (blue/cyan).

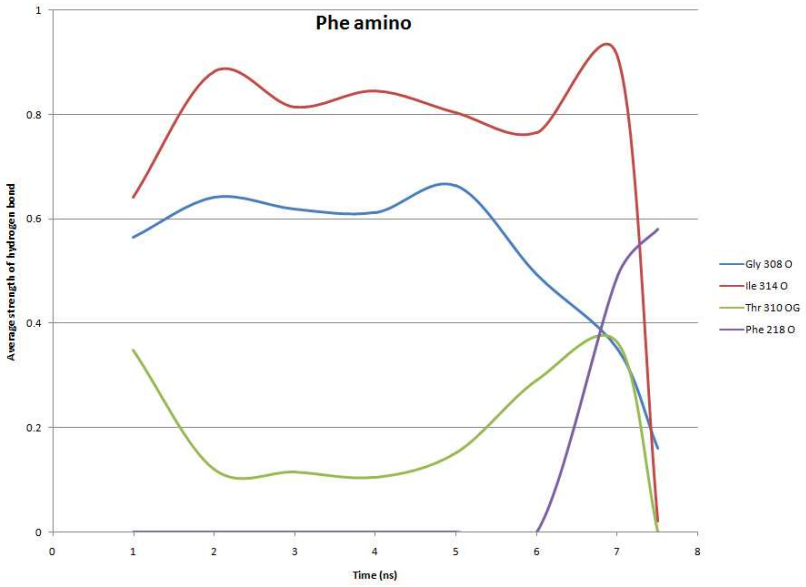


Figure 6.13: **Hydrogen bonding between the L-Phe substrate amino group and PheA in the PheA-Phe-Lys simulation.** Hydrogen bonding is displayed as a measure of strength over time (ns).

was calculated and will be used as a measure of the hydrogen bonding strength between particular residue groups.

No hydrogen bonding interactions are observed in the PheA-Phe-Lys simulation between the Asp 219 carboxyl group and Phe substrate α-amino group, figure 6.13. Strong bonding interactions are however observed between the α-amino group and the main chain carbonyl of Ile 314 during the first seven nanoseconds of the simulation. During the final 500 ps of the simulation hydrogen bonding between these groups is absent. Fairly strong (0.6) hydrogen bonding interactions are formed between the Thr 310 hydroxyl side chain and the Phe α-amino group. The strength of this bond diminishes towards the end of the simulation. From 6 ns onwards a hydrogen bonding interaction of increasing strength is observed between the Phe substrate α-amino group and Phe 219 main chain carbonyl group.

Initially strong hydrogen bonding interactions are observed in the PheA-Phe-Lys simulation between the Lys 501 amino group and Phe substrate α-carboxyl group, figure 6.14. During the sixth nanosecond this interaction weakens and a stronger interaction is formed between the hydroxyl side chain of Ser 498 and α-carboxyl group of the Phe substrate. From 2 ns onwards one hydrogen bond is formed between the ND2 side chain of Asn 499 and the Phe substrate α-carboxyl group.

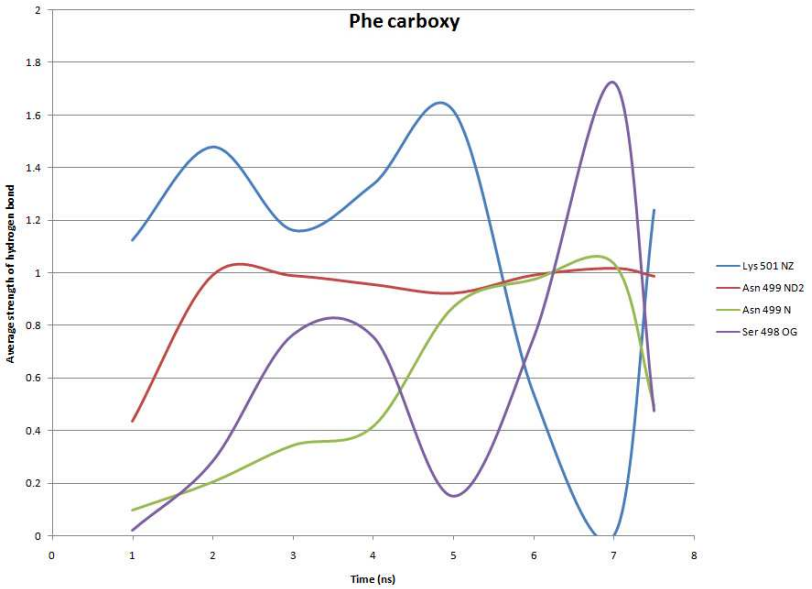


Figure 6.14: **Hydrogen bonding between the L-Phe substrate carboxyl group and PheA in the PheA-Phe-Lys simulation.** Hydrogen bonding is displayed as a measure of strength over time (ns).

As is the case in the PheA-Phe-Lys simulation, no hydrogen bonding interactions are observed in the PheA-Asp-Lys simulation between the Asp 219 α-carboxyl group and Asp substrate amino group, figure 6.15. The strong bonding between the amino group and the main chain carbonyl of Ile 314 O observed during the first seven nanoseconds of the PheA-Phe-Lys simulation is not present in the PheA-Asp-Lys simulation. The fairly strong (0.6) hydrogen bonding observed between the Thr 310 hydroxyl side chain and the substrate amino group in the PheA-Phe-Lys simulation is only seen in the seventh nanosecond of the PheA-Asp-Lys simulation. Hydrogen bonding of fairly consistent strength (0.6–1.0) between the α-amino group of the Asp substrate and the main chain carbonyl of Gly 308 is present in PheA-Asp-Lys for the duration of the simulation.

Initially strong hydrogen bonding is observed in the PheA-Asp-Lys simulation between the Lys 501 amino group and Asp substrate α-carboxyl group, figure 6.16. During the fourth nanosecond this interaction weakens and by the sixth nanosecond no hydrogen bonding interactions are formed by this residue, or any other PheA residue and the α-carboxyl group of PheA-Asp-Lys.

The Thr to Asp (262) residue mutation was introduced to PheA in order to aid binding of the Asp substrate. Figure 6.17 shows that on the timescale of the simulation strong

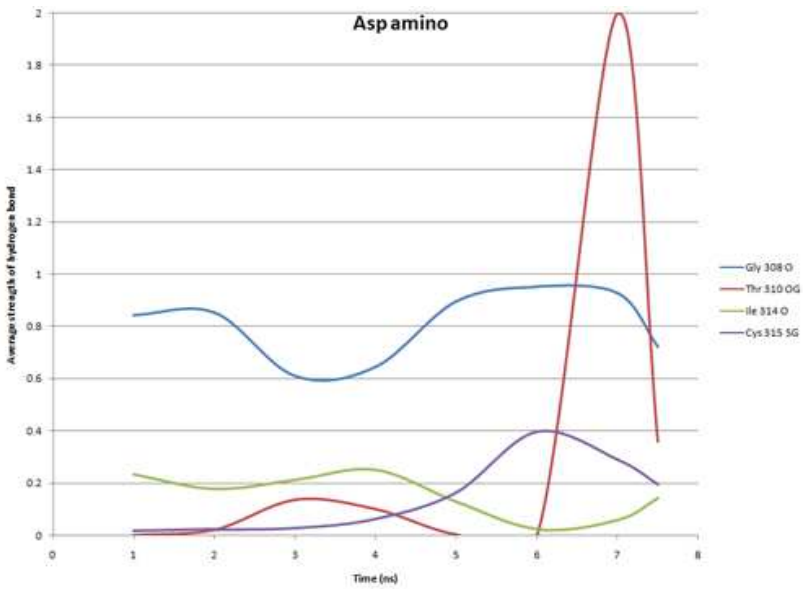


Figure 6.15: **Hydrogen bonding between the L-Asp substrate amino group and PheA in the PheA-Asp-Lys simulation.** Hydrogen bonding is displayed as a measure of strength over time (ns).

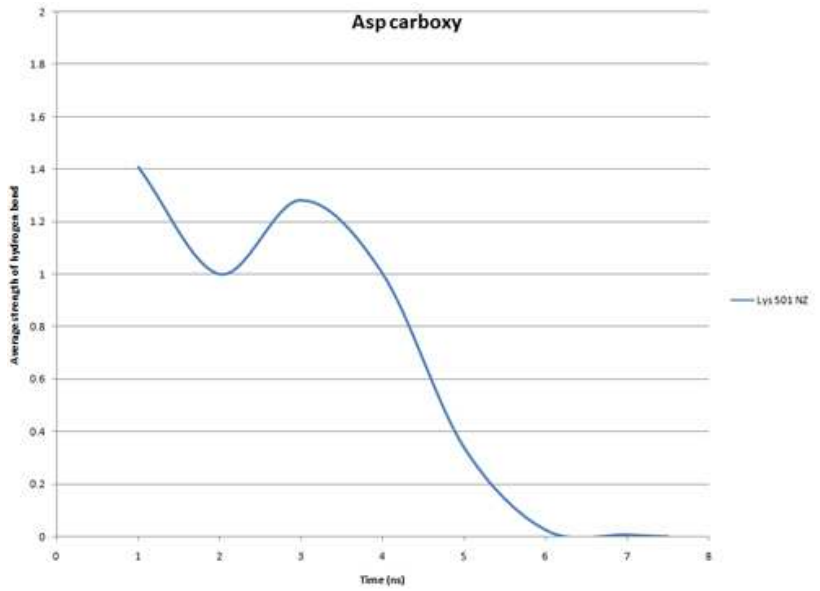


Figure 6.16: **Hydrogen bonding between the L-Asp substrate carboxyl group and PheA in the PheA-Asp-Lys simulation.** Hydrogen bonding is displayed as a measure of strength over time (ns).

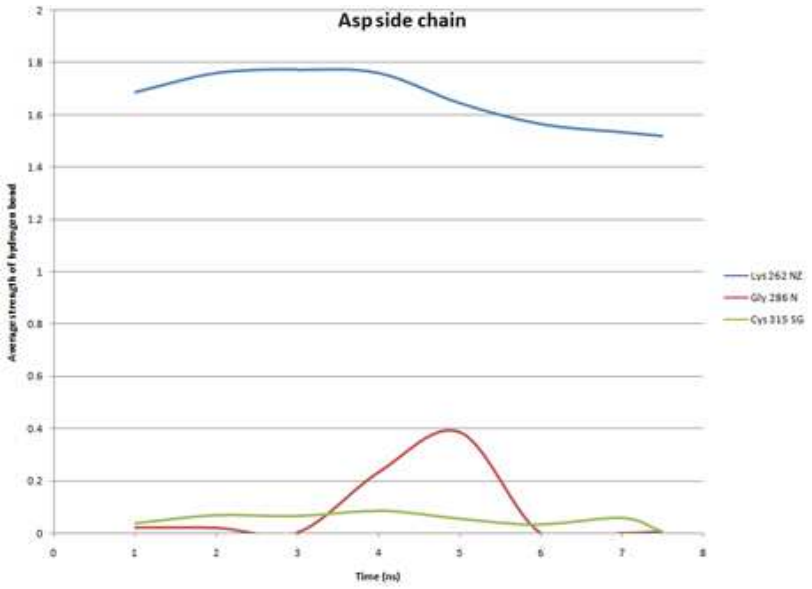


Figure 6.17: **Hydrogen bonding between the L-Asp substrate side chain group and PheA in the PheA-Asp-Lys simulation.** Hydrogen bonding is displayed as a measure of strength over time (ns).

hydrogen bonding is observed between the side chain group of the Asp substrate and the ND2 side chain group of Lys 262. The key interactions identified between the α-amino and α-carboxyl groups of the substrate and Asp 219 and Lys 501 residues are however absent and weak, respectively, in both this simulation and the PheA-Phe-Lys simulation.

The hydrogen bonding of the Phe substrate to the PheA-Phe-His protein, figure 6.18 is mediated by the same groups of PheA that form hydrogen bonds to the Phe substrate in PheA-Phe-Lys. The strength of these hydrogen bonding interactions differ in the two simulations. In PheA-Phe-His strong hydrogen bonding is formed between the Phe α-amino group and the main chain carbonyl of Ile 314. This hydrogen bonding is present throughout the simulation. The initially strong hydrogen bonding between the main chain carbonyl of Gly 308 and Phe substrate amino group weakens after the fourth nanosecond and is intermittently present for the remainder of the simulation. Hydrogen bonding of increasing strength is formed between the Phe substrate α-amino group and Thr 310 hydroxyl side chain.

The hydrogen bonding interaction between the α-carboxyl group of the Phe substrate in the PheA-Phe-His simulation, see figure 6.19, is of comparable magnitude and duration to the equivalent bond formed in the PheA-Asp-Lys simulation. As is the case in all of the

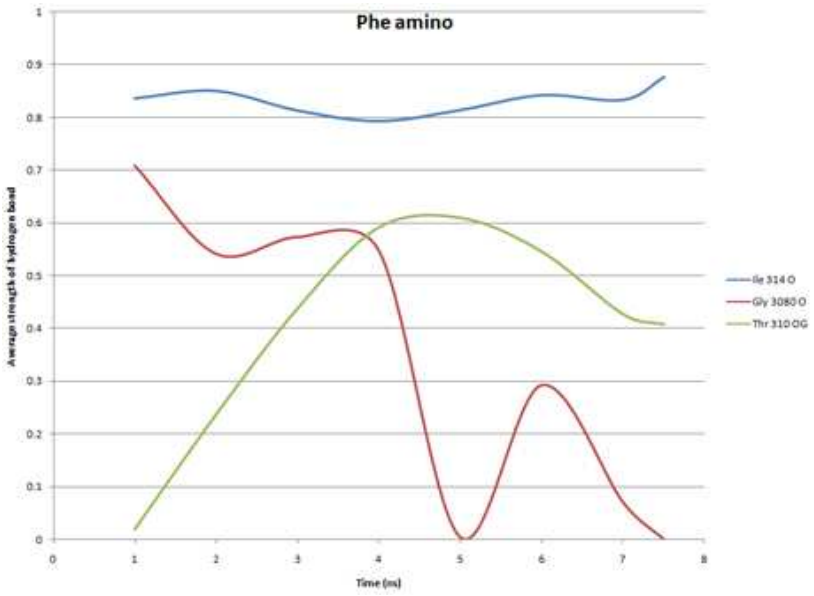


Figure 6.18: **Hydrogen bonding between the L-Phe substrate amino group and PheA in the PheA-Phe-His simulation.** Hydrogen bonding is displayed as a measure of strength over time (ns).

simulations with mutated PheA, no hydrogen bonding is observed between the α-amino group of the Asp substrate and Asp 219 in the PheA-Asp-His simulation, see figure 6.20. Hydrogen bonding is however observed between the main chain carbonyl group of Gly 308 and the Asp substrate.

Additionally no hydrogen bonding is observed in the PheA-Asp-His simulation between the Asp substrate carboxyl group and Lys 501 amino side chain group, see figure 6.21. Instead interaction of the carboxyl group of the Asp substrate is mediated by hydrogen bonding interactions between the side chain hydroxyl group of Thr 174, main chain amino group of Asp 219, main chain amino group of Thr 310 and side chain hydroxyl group of Thr 310.

The Ala to His (306) residue mutation was introduced to PheA in order to aid binding of the Asp substrate. Figure 6.22 shows that on the timescale of the simulation no hydrogen bonding is observed between this PheA residue and the Asp substrate. Hydrogen bonding is however briefly observed between the Asp substrate side chain and Lys 501 amino group (2–4 ns), and the Asp substrate side chain and side chain hydroxyl of Ser 175 of the A3 loop.

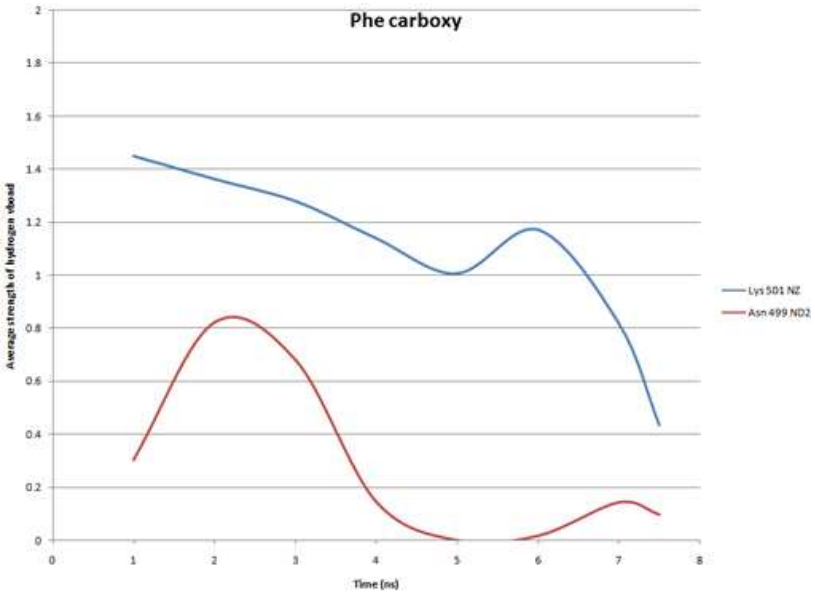


Figure 6.19: **Hydrogen bonding between the L-Phe substrate carboxyl group and PheA in the PheA-Phe-His simulation.** Hydrogen bonding is displayed as a measure of strength over time (ns).

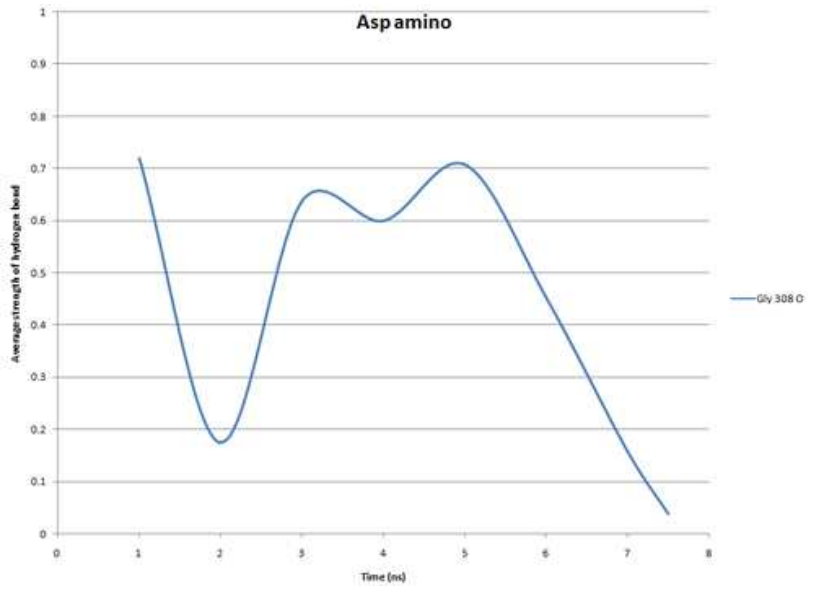


Figure 6.20: **Hydrogen bonding between the L-Asp substrate amino group and PheA in the PheA-Asp-His simulation.** Hydrogen bonding is displayed as a measure of strength over time (ns).

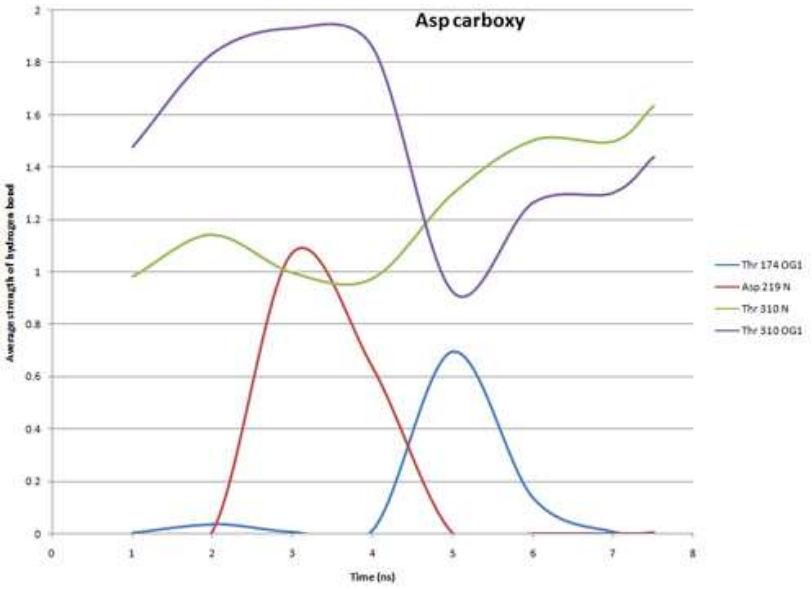


Figure 6.21: **Hydrogen bonding between the L-Asp substrate carboxyl group and PheA in the PheA-Asp-His simulation.** Hydrogen bonding is displayed as a measure of strength over time (ns).

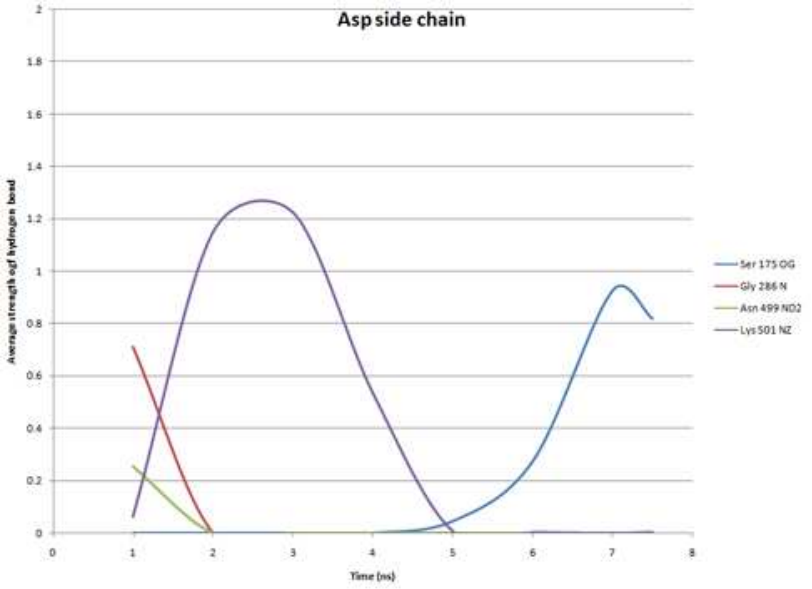


Figure 6.22: **Hydrogen bonding between the L-Asp substrate side chain group and PheA in the PheA-Asp-His simulation.** Hydrogen bonding is displayed as a measure of strength over time (ns).

6.2.6 Summary

The key interactions formed between the Phe substrate and PheA protein identified in the holo simulations of PheA are absent or weak in all of the PheA mutation simulations. The mutated residue 262 (Thr to Lys) does form hydrogen bonding interactions of PheA to the side chain of the Asp substrate. Interdomain motion of the A_{sub} domain of PheA relative to the A_{core} domain of PheA is only seen in the PheA-Phe-Lys simulation. This motion, towards the right side of PheA away from the A3 motif, is similar to that observed in the PheA-apo simulations.

6.3 Metdynamics Calculations - Set Up

To build on the observations from the classical MD simulations of apo and holo PheA presented in Chapter 3, free energy calculations, using the metadynamics method have been designed, parameterised and set up. These calculations have been designed to characterise the free energy landscape of the observed PheA motion and explore whether the PheA enzyme is capable of the full large scale domain rotation required to adopt the second forming conformation observed in other members of the superfamily.

Each calculation has been designed to determine the free energy profile from the PheA enzyme in the first half reaction conformation to the second half reaction conformation; the second half reaction conformation of the PheA enzyme has been obtained by comparative modelling with DltA, a D-alanine D-alanyl carrier protein ligase. Two parallel sets of calculations have been set up, each with a different intermediate state. The two possible intermediate structures along the path have been taken from the extreme projection along the trajectory of the first eigenvector of motion from each of the earlier PheA-holo MD simulations. These calculations are being performed in the apo state of the protein to explore whether the PheA enzyme is capable of the "domain alternation" rotation and if this rotation differs in the presence and absence of the first half reaction hydrolysed products. The collective variable has been selected as the RMSD of the A8 motif residues of the PheA

which span the region between the domains.

Chapter 7

Conclusion

7.1 Conclusions

7.1.1 Summary of Project

One strategy for producing novel antibiotics and molecules with pharmacologically attractive properties is the modification of the assembly line machinery of the nonribosomal peptide synthetase (NRPS). Although the A domains have been studied extensively, knowledge of the selectivity mechanism is still relatively rudimentary. Understanding the molecular basis of this selectivity is critical for informed reprogramming of these domains. Very little data are currently available that conclusively explains how the individual NRPS domains are oriented with respect to each other and how they interact. The interactions which take place across the domains require a degree of flexibility in the PCP domain and, very likely, in other NRPS domains. In the case of the A domain the substrate and ATP-Mg cofactor enter the active site and react to form a high energy acyl-adenylate (half-reaction 1). Following the release of PP_i the substrate is covalently tethered to the terminal thiol of the Ppant arm of the PCP domain (half-reaction 2).

As previously discussed in section 3.1 of chapter 1 alternation between two conformations, essentially characterised by rotation of the A_{sub} domain, has been proposed as a strategy to reconfigure the enzyme's single active for the catalysis of these two distinct half-reactions [52,53,91]. This strategy has been proposed following observation of members of the adenylate-forming superfamily crystallised with the first and second-half reaction products/substrates.

In this thesis MD simulations of PheA in the apo form, with the cognate substrate (Chapter 3), and noncognate substrates (Chapter 4) are presented. A homology model of the second A domain of the NRPS that forms Coelichelin was built and MD simulations of this A domain in the apo state, with the cognate substrate, and noncogate substrates carried out (Chapter 5). The conclusions from each study are presented in full at the end of the relevant chapter.

While such a large scale domain rotation suggested to reconfigure the single active site in adenylate-forming superfamily would not be observed on the timescale of the simulations

presented in this thesis, in each of the PheA apo and cognate substrate simulations, rotation between the A_{core} domain and A_{sub} or part of the A_{sub} domain was observed.

The molecular modelling study of PheA reveals interesting differences between the inter-domain motion in the PheA Phe substrate apo and holo state simulations. In the apo simulations rotation of the A_{sub} domain is primarily towards the right side of the PheA A_{core} domain away from the A3 motif loop. In contrast the principal motion in the PheA-holo cognate substrate simulations is the rotation of the A_{sub} domain or part of the A_{sub} domain towards the A3 motif loop on the left side of PheA. This motion widens an opening through which the PPant arm may use to access the active site. In each of these rotations residues from the conserved A8 motif loop were identified as flexible and to act as hinge residues. These hinge residues include the conserved Asp residue (414, pdb: 430) and other residues that form the loop that follows subdomain D of the A_{sub} domain.

The two holo cognate substrate PheA simulations revealed the critical nature of the highly conserved Asp and invariant Lys residues for holding the substrate in a productive confirmation and the results from these simulations suggest these key interactions are required for the domain rotation observed. Hydrogen bonding between the Asp 219 (pdb:235) residue and a residue from the A3 loop (Thr 174, pdb: 190) in the PheA2-holo simulation suggests a role for the A3 loop in stabilising the enzyme to maintain the opening between the domains through which the PPant may access the active site of the enzyme or this interaction may be an intermediate stabilising interaction required to facilitate further rotation of the A_{sub} domain. This observation has not been previously postulated in the literature. This interaction is not observed in the PheA1-holo simulation where the motion described by the second eigenvector is different to that described by the second eigenvector of the PheA2-holo simulation. The difference in motion between two cognate substrate holo simulations is not unusual as in general a trajectory samples only one region and few transitions between regions are observed. Additional simulations would provide greater sampling of the phase space.

The simulation with the noncognate L-Asp substrate, presented in Chapter 4, additionally suggests the A3 motif loop may have a role in removing noncognate ligands from the en-

zyme's active site. The L-Asp substrate leaves the PheA binding pocket on the timescale of the simulation most likely as a result of the lack of suitably positioned residues in the binding pocket for the L-Asp sidechain to form hydrogen bonding interactions with. This observation is consistent with the observations of Ackerley and co-workers and Lautru and co-workers who suggest, from analysis of the substrate specificity code, that smaller substrates are thought to utilise only the residues at the top of the biding pocket [30,85]. As the L-Asp substrate leaves the binding pocket it forms hydrogen bonding interactions with residues from the A3 motif loop. Domain rotation is observed in the PheA-Tyr and analysis of the hydrogen bonding interactions between the substrate and PheA in this simulation supports the suggestion that interactions between the substrate and Asp and Lys binding pocket residues are necessary for the interdomain rotation.

The simulations with the homology model revealed that the core regions of structure are stable on the timescale of the simulations. Domain rotation was observed in the CchH2-Thr and CchH2-Ser simulations. As assessment of substrate binding was made based on the hydrogen bonding interactions between the substrates and CchH2. This analysis showed the L-Thr specific CchH2 A domain binds the L-Thr substrate well and better than non-cognate substrates. The interaction between the substrate and key binding pocket residues in the CchH-Thr simulation and initial stages of the CchH2-Ser and the corresponding domain rotation observed supports the suggestion that the these key interactions are required for the domain rotation which widens the opening on the right side of the protein through which the PPant arm may access the active site. The results from this study suggest that homology modelling of the A domains may be a useful technique for further study of the dynamics and substrate interactions of the A domains.

In summary, the work presented in this thesis provides evidence that rotation of the A_{sub} domain occurs in the PheA A domain in the presence of the hydrolysed products of the first half reaction. This rotation widens an opening between the domains through which the PPant arm may access the active site to carry out the second half reaction. A key residue (Thr 174, pdb: 190) from the A3 motif loop has been identified as forming an interaction with one of the conserved key binding pocket residues. This interaction may

served to stabilise the enzyme to maintain the opening between the domains through which the PPant may access the active site of the enzyme or it may be an intermediate stabilising interaction required to facilitate further rotation of the A_{sub} domain. A role for the A3 motif in assisting the removal of noncognate substrates from the binding site has also been suggested from the results of the PheA noncognate L-Asp simulation.

7.1.2 Future Directions

Molecular dynamics can be used to provide insight into the dynamics of proteins at a molecular level. In order for an MD simulation to produce fully meaningful results the run should be long enough for the system to sample all of the energetically relevant configurations[303]. This may not always be possible in practice:

- the system may diffuse very slowly in configuration space, requiring very long simulations at great computational cost
- the relevant configurations may be separated by high free-energy barriers requiring activation of rare fluctuations to take the system over the barriers from one metastable state to another

The large scale domain rotation suggested to reconfigure the single active site in adenylate-forming superfamily would not be observed on the timescale of the simulations presented in this thesis. A number of methods have been developed in recent years to overcome the timescale problem of classical MD. Metadynamics[304] is one such method, others include adaptive force bias and steered MD. Metadynamics enhances sampling through the addition of a bias potential that acts on a specified number of degrees of freedom, referred to as Collective Variables (CVs) and reconstructs the free-energy surface as a function of these CVs. The selection of appropriate CVs that can discriminate between the states of interest is crucial for the quality of results obtained from these calculations.

To build upon the work carried out in this thesis, free energy calculations using the metadynamics method should be performed to characterise the free energy landscape of the motion

observed in the PheA-holo cognate simulations and explore whether the PheA enzyme is capable of the large scale domain rotation required to adopt the thioester forming conformation as observed in other members of the superfamily. The PheA crystal structure and structures from the extremes of motion observed in these simulations could be used as starting structures for these calculations, with the end state determined by modelling the orientation of the A_{sub} domain on one of the structures from the adenylate-forming superfamily in the second half conformation. Preliminary work on setting up metadynamics calculations on the apo state PheA structure has been carried out as described in Chapter 6. Metadynamics calculations with both the hydrolysed products from the first half reaction and the substrates from the first half reaction should be carried out to aid understanding of the role of the A3 motif loop in the domain rotation; the phosphate groups of ATP are proposed to sterically hinder the mobility of the A3 motif loop. Calculations on these structures with mutations of residues in the A3 motif loop could additionally be performed to understand the role of the A3 motif loop residues in the domain rotation.

Appendix I

Domain	Core[a]	Consensus sequence[b]
Adenylation	A1	L(TS)YxEL
	A2 (core 1)	LKAGxAYL(VL)P(LI)D
	A3 (core 2)	LAYxxYSTG(ST)TGxPKG
	A4*	FDxS
	$A5_{aa}$	NxYGPTETTxx
	$A5_{aryl}$[†]	QVxFMAEGLVN
	A6 (core 3)	GELxJGx(VL)ARGYL
	A7 (core 4)	Y(RK)TGDL
	A8 (core 5)	GRxDxQVKIRGxRIELGEIE
	A9	LPxYM(IV)P
	A10	NGK(VL)DR
PCP	P (core 6)	LGG(DH)SL
Condensation	C1	SxAQxR(LM)(WY)xL
	C2	RHExLRTxF
	C3 (His)	MHHxISDG(WV)S
	C4	YxD(FY)AVW
	C5	(IV)GxFVNT(QL)(CA)xR
	C6	(HD)QD(YD)PFE
	C7	RDxSRNPL
Thioesterase	Te	GxSxG
Epimerization	E1	PIQxWF
	E2 (His)	HHxISDG(WV)S
	E3 (race A)	DxLLxAxG
	E4 (race B)	EGHGRE
	E5 (race C)	RTVGWFTxxYP(YV)PFE
	E6	PxxGxGYG
	E7 (race D)	FNYLG(QR)
Cyclization	Cy1	FPL(TS)xxQxAYxxGR
	Cy2	RHx(IM)L(PAL)x(ND)GxQ
	Cy3	LPxxPxLPLxxxP
	Cy4	(TS)(PA)xxx(LAF)x_6(IVT)LxxW
	Cy5	(GA)DFTxLxLL
	Cy6	PVVFTSxL
	Cy7	(ST)(QR)TPQVx(LI)Dx_{13}WD
Oxidation	Ox1	KYxYxSxGxxY(PG)VQ
	Ox2	GxxxG(LV)xxGxYYY(HD)P
	Ox3	IxxxYG
N-Methyltransferase	M1 (SAM)	VL(DE)xGxGxG
	M2	NELSxYRYxAV
	M3	VExSxARQxGxLD
Reductase	R1	V[L][L]TG[A]TG[F][L]GxxLL
	R2	Vx[L][L]VR[A]
	R3	GPL[G]x[P]x[L]GL
	R4	V[Y]PYxYLxx[P]NVxxT
	R5	GYxxSKW[A][A]E
	R6	R[P]G
	R7	Yx_4G(LF)LxxP

Table 7.1: **Conserved motifs of the NRPS domains[3,6,22].** [a]Former nomenclature is given in brackets; [b]Single letter amino code is used for conserved motif residues; alternate residues are given in parentheses; fairly conserved residues in square brackets; x denotes any amino acid; numbers in braces indicate the spacing between conserved residues; * Motif A4 differs in aryl acid activating domains; [†] A5 motif from aryl activating domain[73].

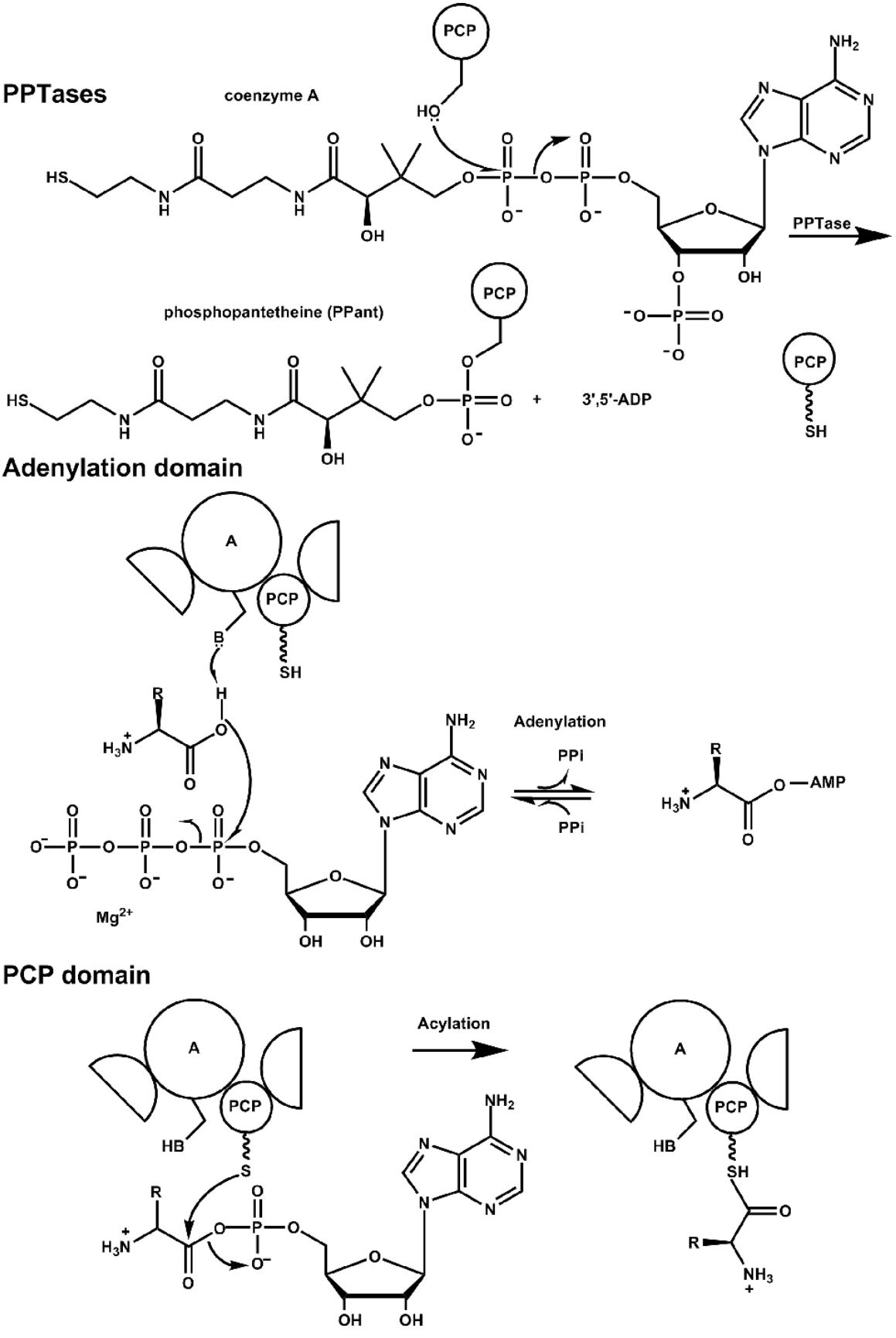


Figure 7.1: **Mechanism of PPTases and A domains**. The mechanism of post-translation modification of PCP domains by PPTases, amino acid adenylation and acylation by the A domain. Modified from figures 14 and 23 of[305].

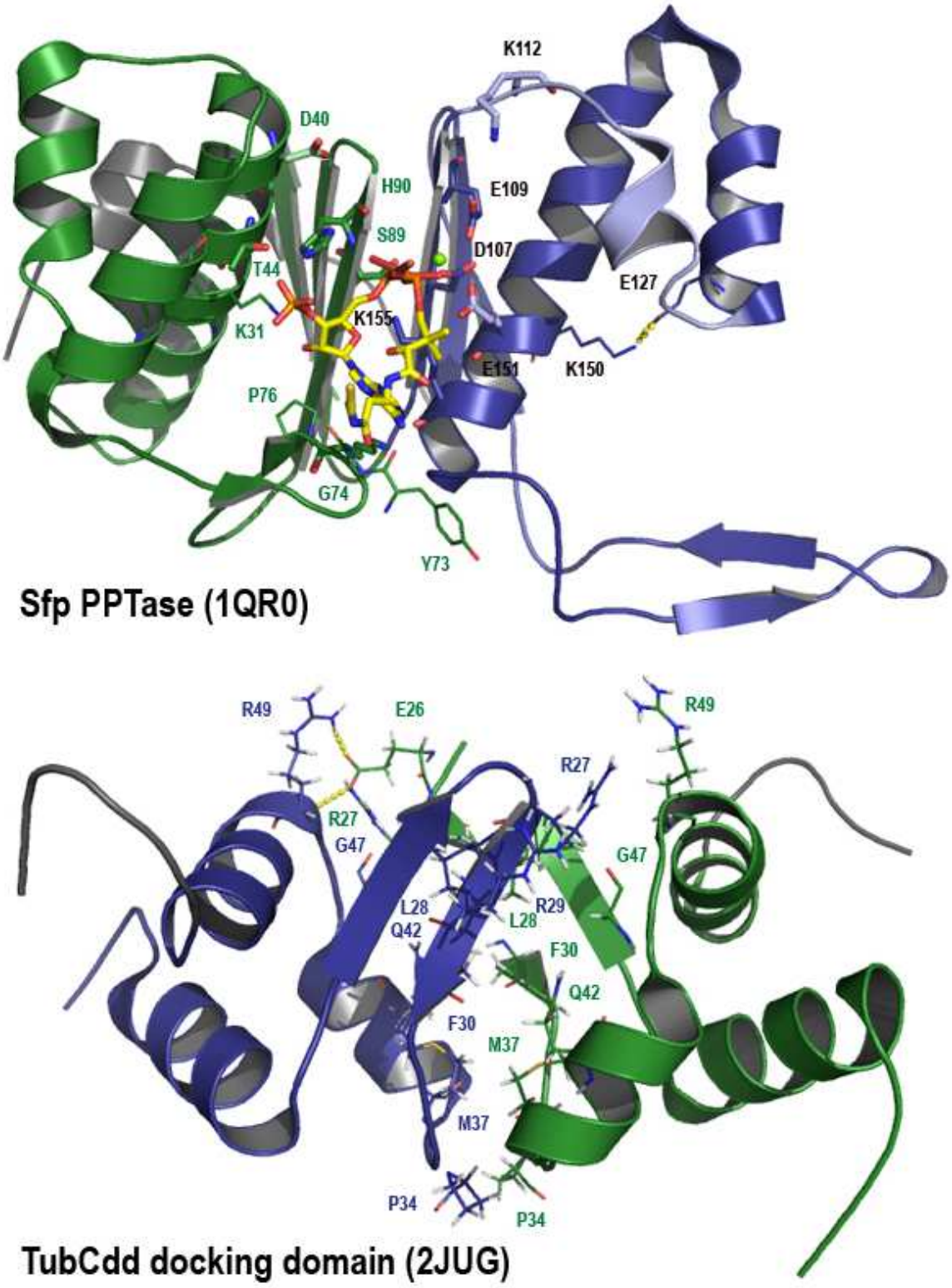


Figure 7.2: **The structure of Sfp and of TubCdd**. PPTase Sfp[306] and the hybrid megasynthetase docking domain TubCdd dimer[307].

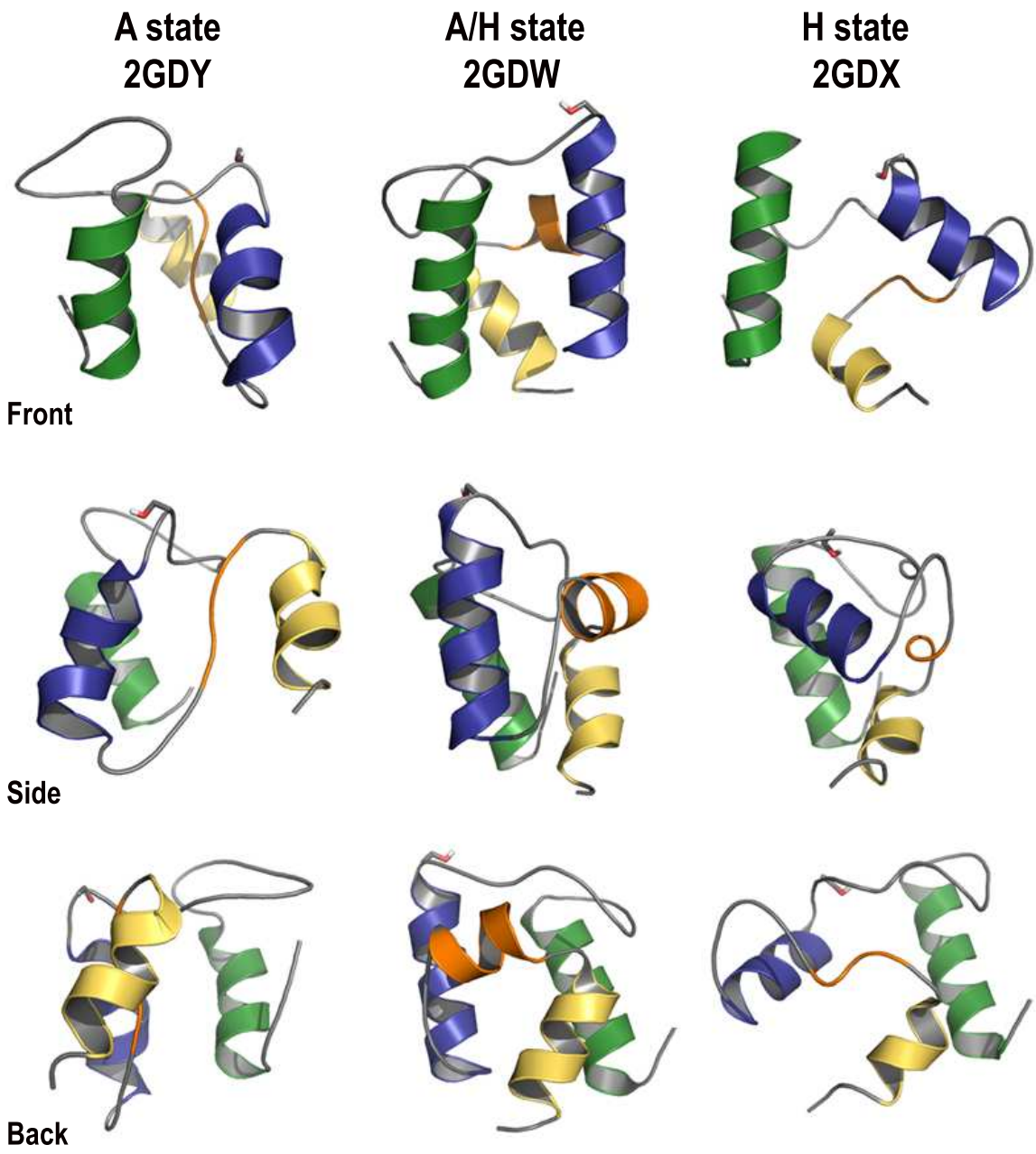


Figure 7.3: **The average NMR solution structures of the TycC3-PCP conformers** in the A, A/H and H states viewed from three alternate angles[141].

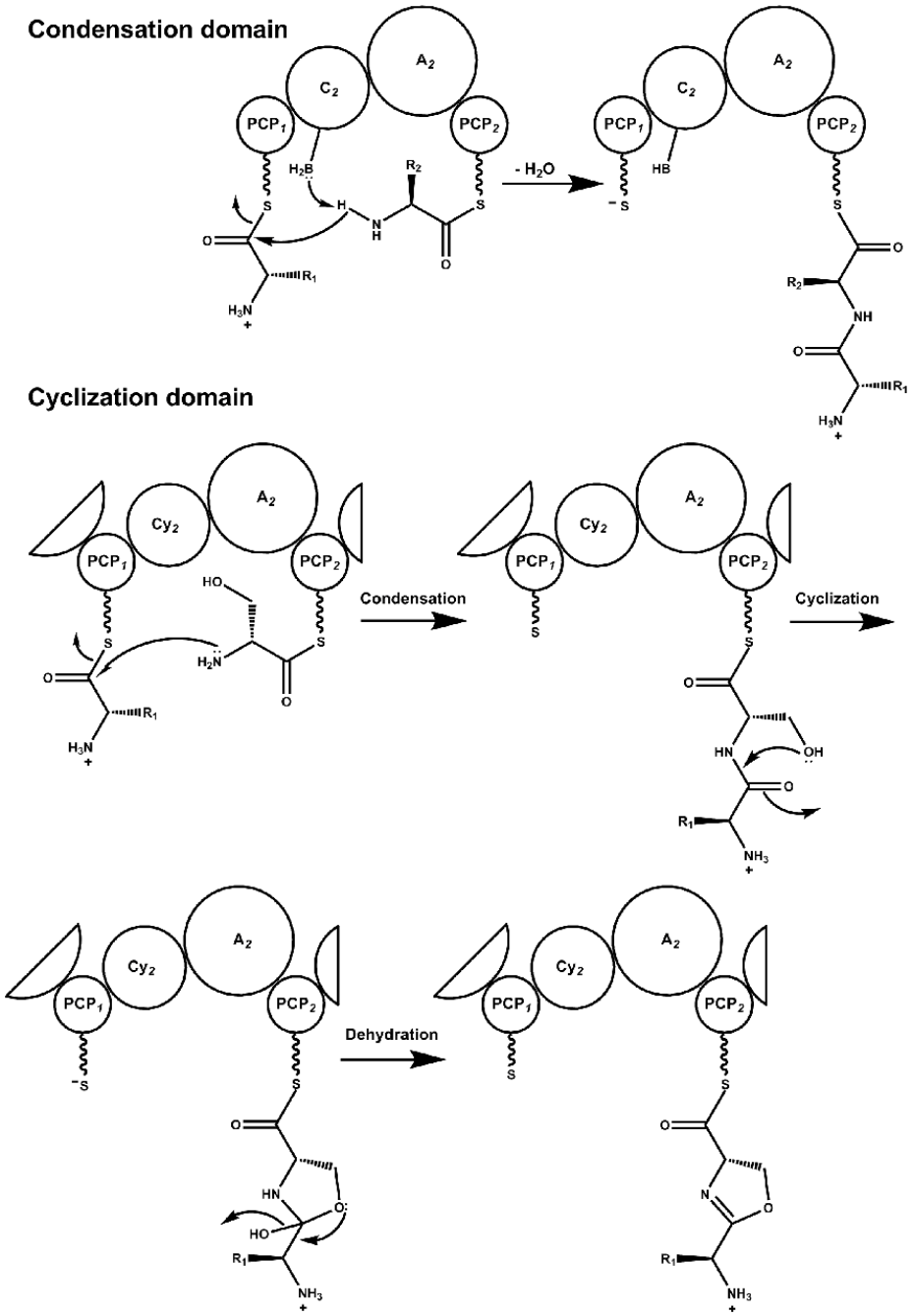


Figure 7.4: **Mechanism of condensation and cyclization**The mechanism by which the C domain catalyzes peptide (C-N) bond formation between the electrophilic upstream peptidyl-*S*-T_1 and the nucleophilic downstream aminoacyl-*S*-T_2. Cy domains catalyze condensation of the peptide bond and then catalyze the attack of the β-nucleophile on the upstream amide carbonyl. This forms a five-membered adduct that dehydrates either to an oxazoline or a thiazoline. Modified from figure 23 of[305] and figure 21 of[308].

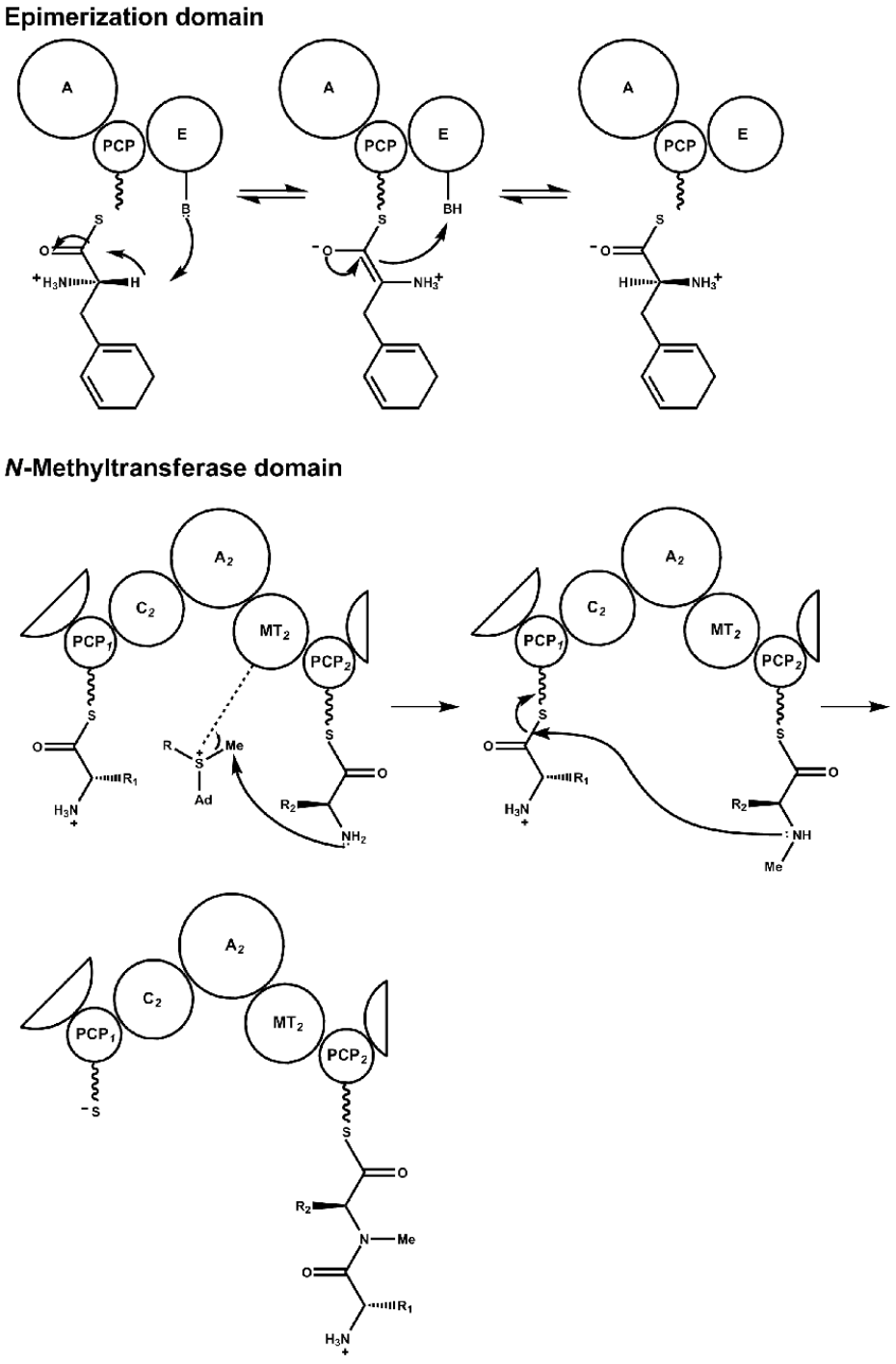


Figure 7.5: **Mechanism of epimerization and methylation** The E domain epimerizes the upstream L-peptidyl-*S*-T acyl donor to D-peptidyl-*S*-T prior. This occurs prior to condensation and the downstream C domain is D-peptidyl-*S*-T specific ensuring that condensation does not precede epimerization. N-Mt domains catalyze transfer of the CH_3 group from *S*-adenosylmethionine (SAM) to the amino group of the aminoacyl-*S*-T intermediate. Figure modified from figures 2 and 1, respectively, from[309].

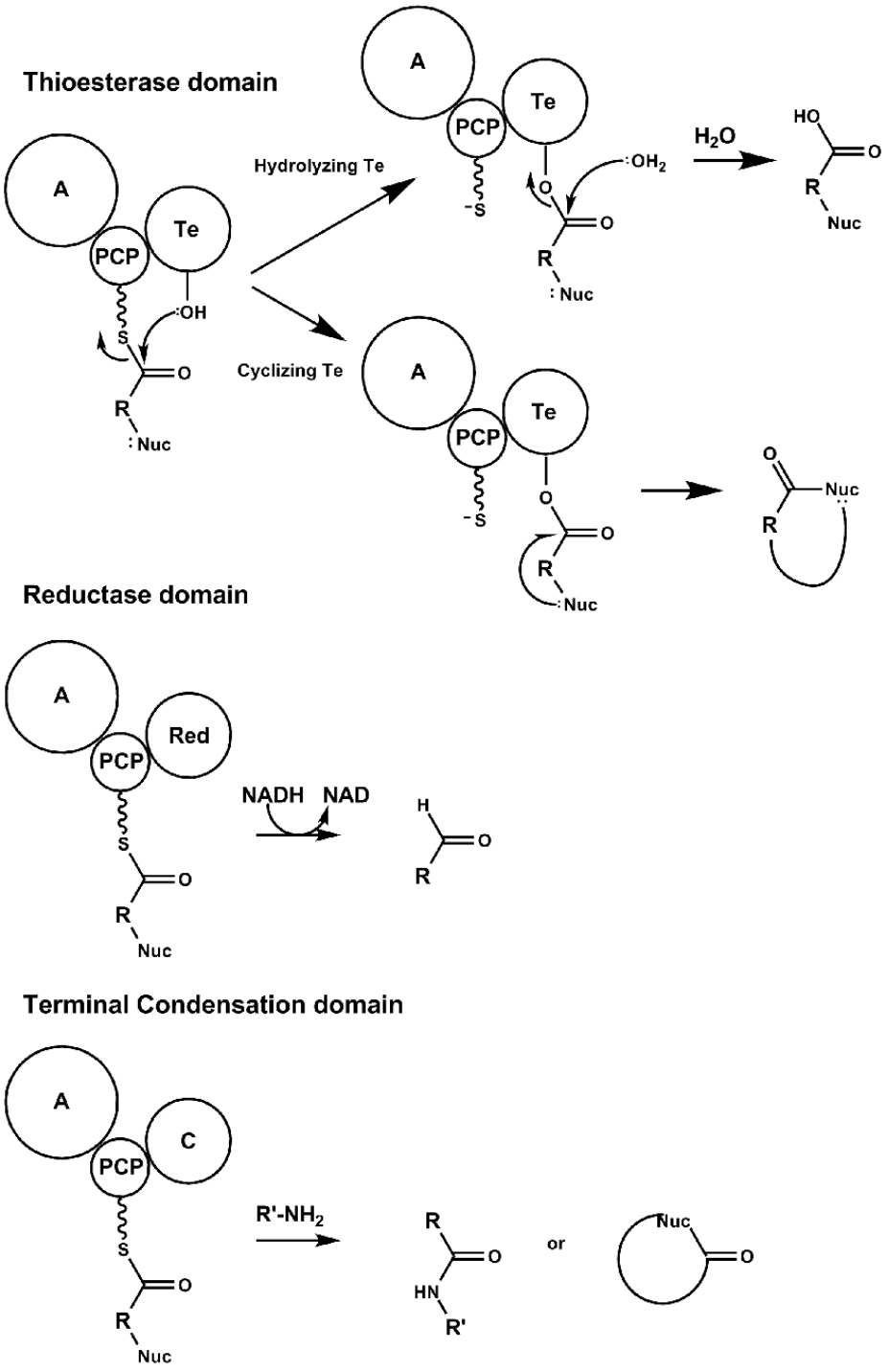


Figure 7.6: **Three strategies for chain termination in NRPSs.** The terminal domain can be a Te, Red or a C domain. Te domains hydrolyze or cyclize the mature chain, NAD(P)H coupled reduction by the Red domain releases an aldehyde, and the scissle thioester bond can be attacked by C domains using either an inter- or intramolecular nucleophile. Image adapted from figure 3 from[310].

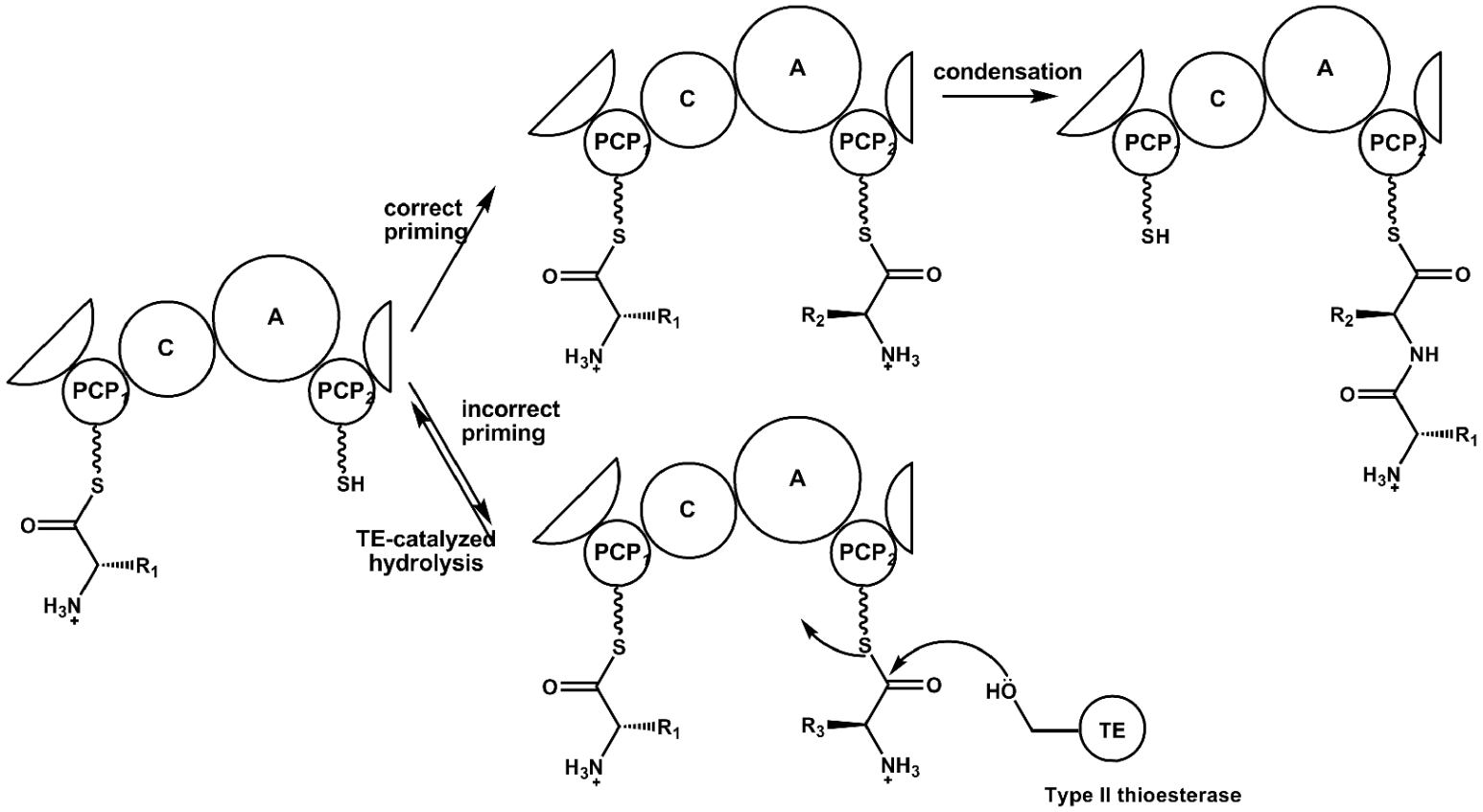


Figure 7.7: **Mechanism of type II thioesterase action**Type II TE domains reverse mispriming of the PCP domains. This regenerates the free pantetheinyl thiol and prevents stalling of the assembly line. Image adapted from figure 34 of[305].

Appendix II

[atoms]

nr	type	resnr	residue	atom	cgnr	charge	mass
1	NR	1	AMP	N9	1	-0.200	14.0067
2	C	1	AMP	C4	1	0.200	12.0110
3	NR	1	AMP	N3	2	-0.360	14.0067
4	C	1	AMP	C2	2	0.220	12.0110
5	HC	1	AMP	H2	2	0.140	1.0080
6	NR	1	AMP	N1	3	-0.360	14.0067
7	C	1	AMP	C6	3	0.360	12.0110
8	NT	1	AMP	N6	4	-0.830	14.0067
9	H	1	AMP	H61	4	0.415	1.0080
10	H	1	AMP	H62	4	0.415	1.0080
11	C	1	AMP	C5	5	0.000	12.0110
12	NR	1	AMP	N7	5	-0.360	14.0067
13	C	1	AMP	C8	5	0.220	12.0110
14	HC	1	AMP	H8	5	0.140	1.0080
15	CH1	1	AMP	C1*	6	0.200	13.0190
16	OA	1	AMP	O4*	6	-0.360	15.9994
17	CH1	1	AMP	C4*	6	0.160	13.0190
18	CH1	1	AMP	C2*	7	0.150	13.0190
19	OA	1	AMP	O2*	7	-0.548	15.9994
20	H	1	AMP	H2*	7	0.398	1.0080
21	CH1	1	AMP	C3*	8	0.150	13.0190
22	OA	1	AMP	O3*	8	-0.548	15.9994
23	H	1	AMP	H3*	8	0.398	1.0080
24	CH2	1	AMP	C5*	9	0.000	14.0270
25	OA	1	AMP	O5*	10	-0.520	15.9994
26	P	1	AMP	P	10	1.490	30.9738
27	OM	1	AMP	O1P	10	-0.990	15.9994
28	OM	1	AMP	O2P	10	-0.990	15.9994
29	OM	1	AMP	O3P	10	-0.990	15.9994

Table 7.2: AMP ff43a2 topology file - atoms

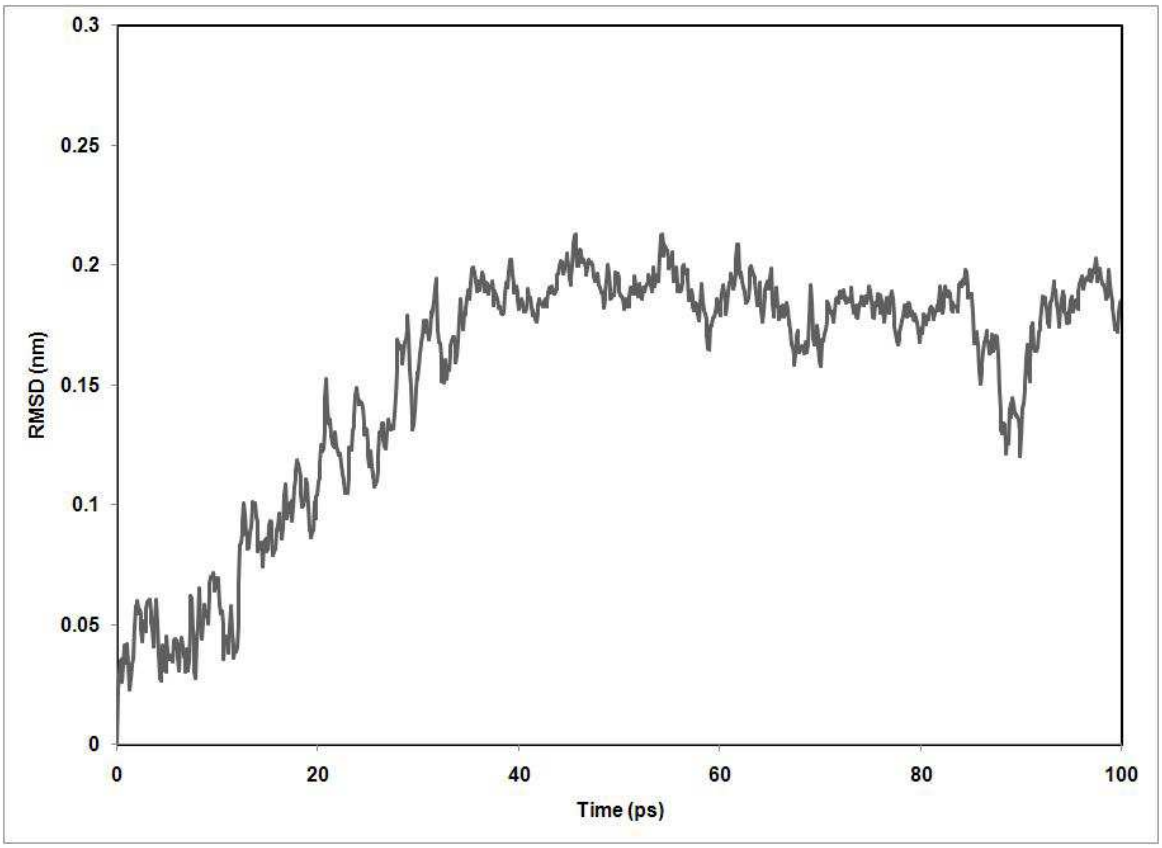
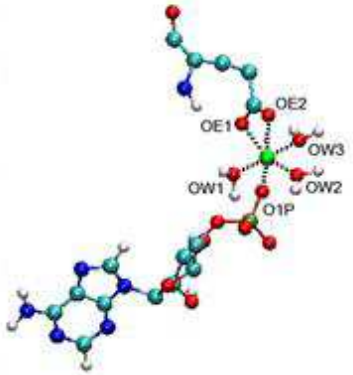


Figure 7.8: **RMSD AMP 1 ns simulation.** The conformational drift of AMP, measured as all heavy atom root mean square deviation (RMSD) from the starting structure. RMSDs vs. time are shown for the AMP heavy atoms (black)

	Bond	Mg-OE1	Mg-OE2	Mg-O1P	Mg-OW1	Mg-OW2	Mg-OW3
Whole simulation	Mean	0.210	0.214	0.197	0.200	0.197	0.201
	SD	0.006	0.008	0.004	0.007	0.006	0.007
First ns	Mean	0.210	0.214	0.198	0.201	0.197	0.200
	SD	0.007	0.010	0.004	0.007	0.006	0.007
Last ns	Mean	0.210	0.214	0.198	0.199	0.196	0.201
	SD	0.006	0.008	0.004	0.006	0.006	0.008

Table 1. Mg coordination distances

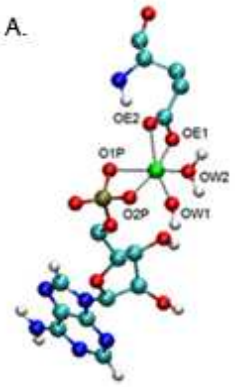


	Angle	OE1-OE2	OE1-OW2	OE2-O1P	OE1-OW3	OE2-OW3	OW3-O1P	OW2-OW3	O1P-OW2	OW1-OE2	OW1-OW2	OW1-OE1	OW1-O1P
Whole simulation	Mean	60.55	94.19	104.14	90.47	91.07	90.47	88.86	101.15	89.66	89.06	91.71	89.47
	SD	1.81	5.49	6.35	4.28	4.57	3.99	3.95	6.15	4.17	3.97	4.41	4.02
First ns	Mean	60.64	93.97	102.64	91.02	90.93	88.94	88.68	102.77	90.98	88.91	91.89	89.98
	SD	2.11	5.68	6.92	4.29	4.67	3.73	4.05	7.86	4.40	3.91	4.30	4.26
Last ns	Mean	60.45	93.02	105.31	90.73	92.00	90.55	89.09	101.22	89.81	89.63	91.21	88.09
	SD	1.72	4.93	6.22	4.29	4.48	3.88	3.79	5.43	4.17	3.96	4.32	3.69

Table 2. Mg coordination distances

Figure 7.9: **Mg ion coordination in the PheA1-holo system simulation.** The distances (table 1) and angles (table 2) between the magnesium ion and the six ligands in the distorted octahedral geometry; OE1 and OE2 from Glu 311, O1P from AMP and three water oxygen atoms.

A.



	Bond	Mg-OE1	Mg-OE2	Mg-O1P	Mg-O2P	Mg-OW1	Mg-OW2
Whole simulation	Mean	0.206	0.210	0.200	0.202	0.196	0.196
	SD	0.006	0.007	0.005	0.005	0.006	0.006
First ns	Mean	0.205	0.210	0.200	0.201	0.196	0.196
	SD	0.005	0.007	0.005	0.005	0.006	0.006
Last ns	Mean	0.206	0.209	0.200	0.202	0.196	0.197
	SD	0.006	0.006	0.005	0.005	0.005	0.006

Table 1. Mg coordination distances

	Angle	OE1-OE2	OE1-OW2	OE1-O2P	OE1-O1P	OE2-OW2	OW2-O2P	O2P-O1P	O1P-OE2	OW1-OE2	OW1-OW2	OW1-O2P	OW1-O1P
Whole simulation	Mean	61.95	104.04	103.33	93.97	90.06	93.89	66.61	90.71	95.48	90.46	100.31	96.99
	SD	1.68	7.61	7.07	6.57	4.47	4.74	1.78	5.87	5.38	4.20	5.89	5.74
First ns	Mean	62.01	105.15	102.27	93.32	90.04	94.13	66.44	93.13	95.41	89.54	101.74	95.36
	SD	1.63	7.15	6.62	6.27	4.40	4.37	1.77	5.87	5.12	3.92	5.47	4.86
Last ns	Mean	62.17	102.98	103.88	93.88	89.94	95.10	66.85	91.73	95.64	90.17	99.10	95.42
	SD	1.70	6.30	5.87	5.73	4.36	4.86	1.65	5.38	5.46	3.95	5.50	4.94

Table 2. Mg coordination angles

Figure 7.10: **Mg ion coordination in the PheA2-holo system simulation.**. The distances (table 1) and angles (table 2) between the magnesium ion and the six ligands in the distorted octahedral geometry; OE1 and OE2 from Glu 311, O1P and O2P from AMP and two water oxygen atoms.

[bonds]

ai	aj	funct	
1	2	2	gb_9
1	13	2	gb_9
1	15	2	gb_21
2	3	2	gb_11
2	11	2	gb_15
3	4	2	gb_11
4	5	2	gb_3
4	6	2	gb_11
6	7	2	gb_11
7	8	2	gb_8
7	11	2	gb_15
8	9	2	gb_2
8	10	2	gb_2
11	12	2	gb_9
12	13	2	gb_9
13	14	2	gb_3
15	16	2	gb_19
15	18	2	gb_25
16	17	2	gb_19
17	21	2	gb_25
17	24	2	gb_25
18	19	2	gb_19
18	21	2	gb_25
19	20	2	gb_1
21	22	2	gb_19
22	23	2	gb_1
24	25	2	gb_19
25	26	2	gb_27
26	27	2	gb_23
26	28	2	gb_23
26	29	2	gb_23

Table 7.3: AMP ff43a2 topology file - bonds

[pairs]		
ai aj funct		
1	17	1
1	19	1
1	21	1
2	16	1
2	18	1
6	9	1
6	10	1
8	12	1
9	11	1
10	11	1
13	16	1
13	18	1
15	20	1
15	22	1
15	24	1
16	19	1
16	22	1
16	25	1
17	19	1
17	23	1
17	26	1
18	23	1
18	24	1
19	22	1
20	21	1
21	25	1
22	24	1
24	27	1
24	28	1
24	29	1

Table 7.4: AMP ff43a2 topology file - pairs

[angles]

ai	aj	ak	funct	
2	1	13	2	ga_6
2	1	15	2	ga_36
13	1	15	2	ga_36
1	2	3	2	ga_38
1	2	11	2	ga_6
1	13	14	2	ga_35
12	13	14	2	ga_35
3	2	11	2	ga_26
2	3	4	2	ga_26
3	4	6	2	ga_26
3	4	5	2	ga_24
6	4	5	2	ga_24
4	6	7	2	ga_26
6	7	8	2	ga_26
6	7	11	2	ga_26
8	7	11	2	ga_26
7	8	9	2	ga_22
7	8	10	2	ga_22
9	8	10	2	ga_23
2	11	7	2	ga_26
2	11	12	2	ga_6
7	11	12	2	ga_38
11	12	13	2	ga_6
1	13	12	2	ga_6
1	15	16	2	ga_8
1	15	18	2	ga_8
16	15	18	2	ga_8
15	16	17	2	ga_9
16	17	21	2	ga_8
16	17	24	2	ga_8
21	17	24	2	ga_7
15	18	19	2	ga_8
15	18	21	2	ga_7
19	18	21	2	ga_8
18	19	20	2	ga_11
17	21	18	2	ga_7
17	21	22	2	ga_8
18	21	22	2	ga_8
21	22	23	2	ga_11
17	24	25	2	ga_8
24	25	26	2	ga_25
25	26	27	2	ga_13
25	26	28	2	ga_13
25	26	29	2	ga_13
27	26	28	2	ga_28
27	26	29	2	ga_28
28	26	29	2	ga_28

Table 7.5: AMP ff43a2 topology file - angles

[dihedrals] proper

ai	aj	ak	al	funct	
2	1	15	16	1	gd_6
11	7	8	9	1	gd_4
18	15	16	17	1	gd_14
1	15	18	19	1	gd_7
16	15	18	19	1	gd_8
16	15	18	21	1	gd_17
16	15	18	21	1	gd_7
15	16	17	21	1	gd_14
16	17	21	18	1	gd_7
16	17	21	22	1	gd_8
24	17	21	18	1	gd_17
24	17	21	22	1	gd_7
16	17	24	25	1	gd_8
21	17	24	25	1	gd_17
21	17	24	25	1	gd_7
15	18	19	20	1	gd_12
15	18	21	17	1	gd_17
15	18	21	22	1	gd_7
19	18	21	17	1	gd_7
19	18	21	22	1	gd_8
17	21	22	23	1	gd_12
17	24	25	26	1	gd_14
24	25	26	29	1	gd_11
24	25	26	29	1	gd_9

Table 7.6: AMP ff43a2 topology file - proper dihedrals

[dihedrals] improper

ai	aj	ak	al	funct	
1	13	12	11	2	gi_1
1	2	11	12	2	gi_1
1	13	2	15	2	gi_1
2	3	4	6	2	gi_1
2	11	7	6	2	gi_1
2	1	3	11	2	gi_1
2	12	7	11	2	gi_1
2	1	13	12	2	gi_1
2	11	12	13	2	gi_1
3	2	11	7	2	gi_1
3	4	6	7	2	gi_1
4	3	2	11	2	gi_1
4	6	5	3	2	gi_1
4	6	7	11	2	gi_1
7	10	9	8	2	gi_1
7	11	6	8	2	gi_1
11	2	1	13	2	gi_1
13	12	14	1	2	gi_1
15	1	16	18	2	gi_2
15	21	19	18	2	gi_2
17	16	24	21	2	gi_2
17	22	18	21	2	gi_2

Table 7.7: AMP ff43a2 topology file - improper dihedrals

Appendix III

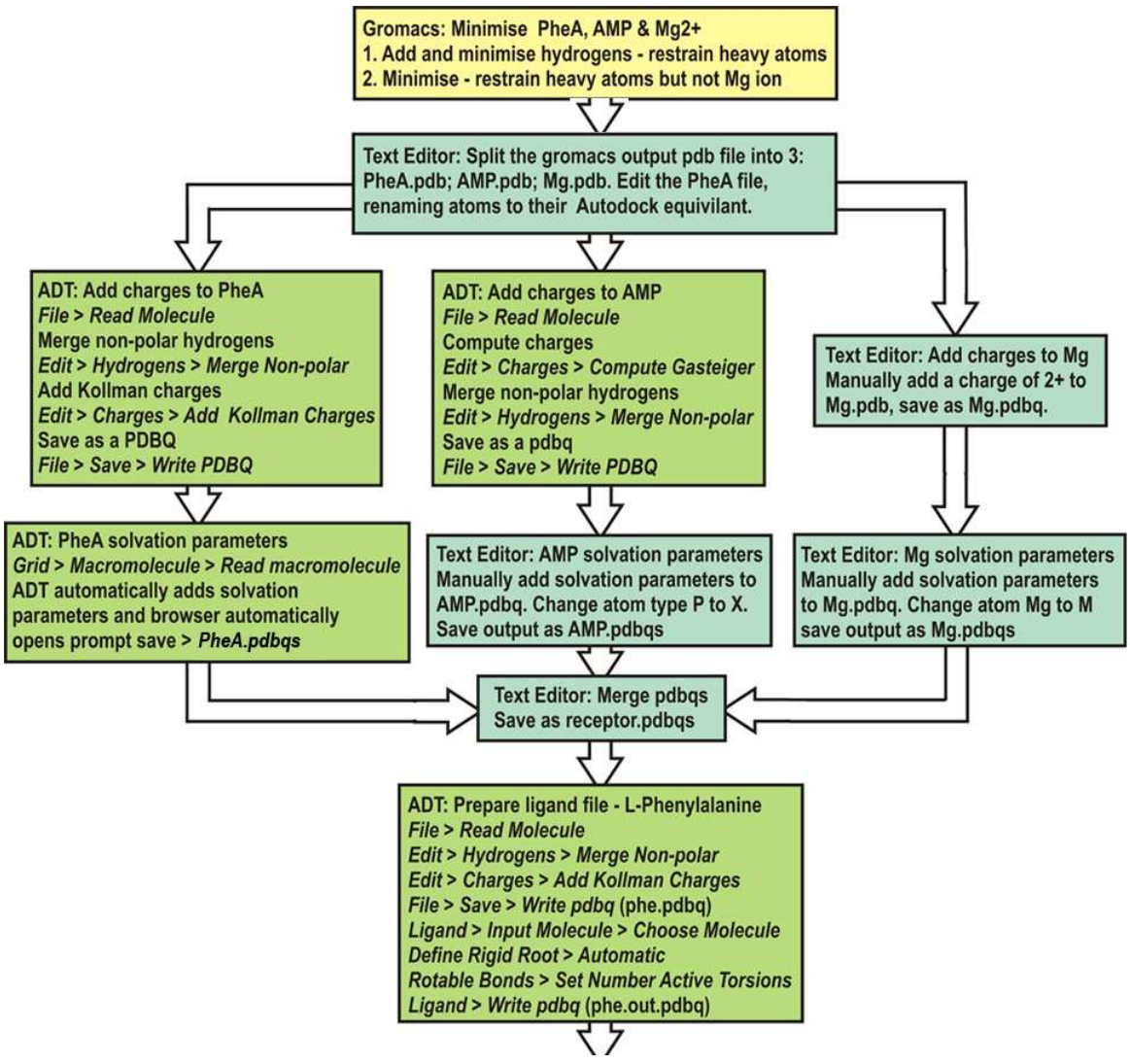


Figure 7.11: Docking flowchart - part 1

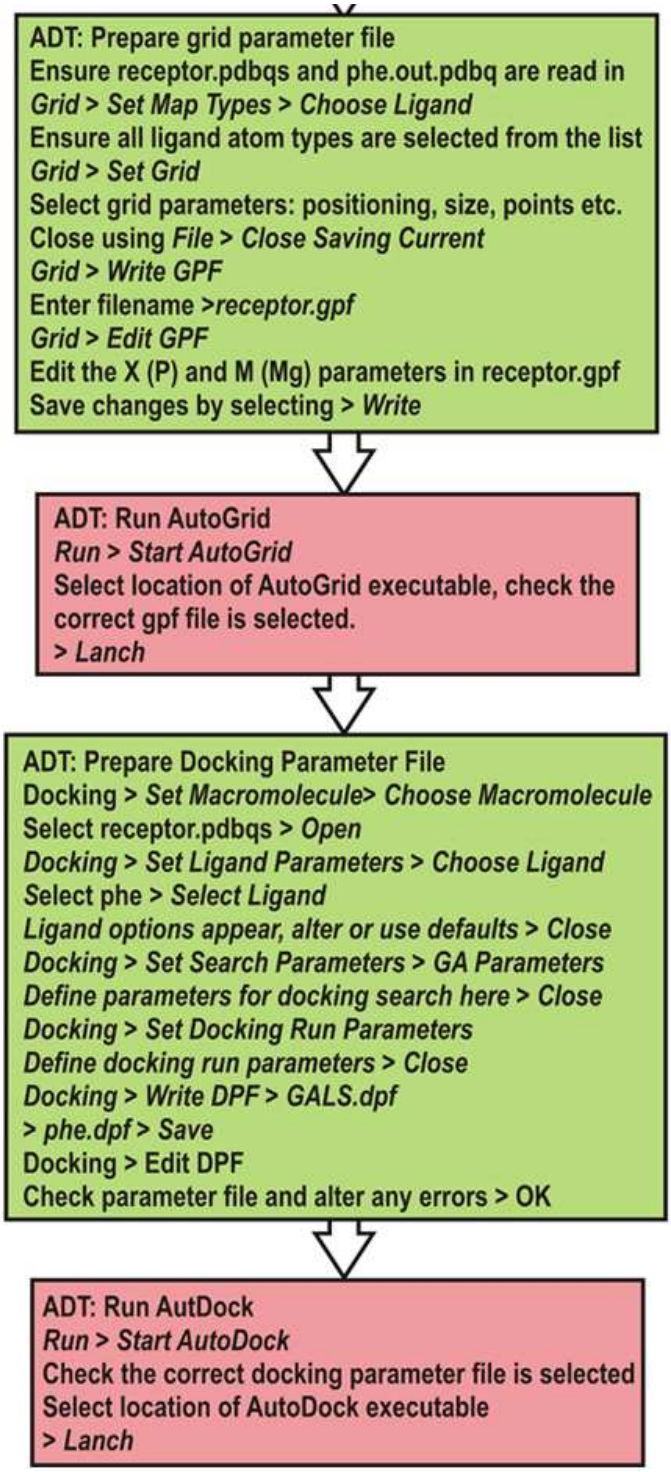
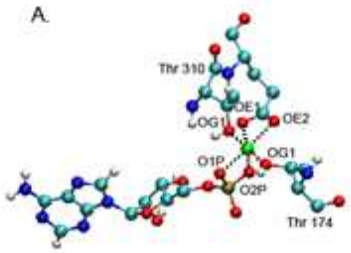


Figure 7.12: Docking flowchart - part 2



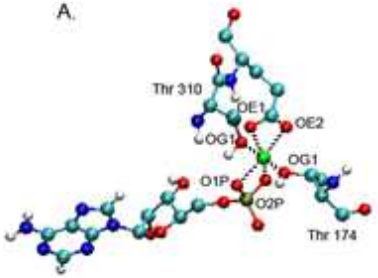
	Bond	Mg-OG1(1)	Mg-OG1(2)	Mg-OE1	Mg-OE2	Mg-O1P	Mg-O2P
Whole simulation	Mean	0.198	0.204	0.207	0.206	0.198	0.199
	SD	0.008	0.009	0.007	0.007	0.005	0.006
First ns	Mean	0.198	0.204	0.206	0.206	0.198	0.198
	SD	0.007	0.009	0.006	0.006	0.004	0.005
Last ns	Mean	1.97	0.205	0.206	0.206	0.197	0.198
	SD	0.013	0.015	0.013	0.012	0.011	0.012

Table 1. Mg coordination distances

	Angle	OE1-OE2	OG1(1)-OG1(2)	OE2-OG1(2)	OE1-O1P	OE2-O2P	OG1(2)-O2P	O2P-O1P	O1P-OE2	OG1(1)-OE1	OG1(2)-OE1	OG1(1)-O2P	OG1(1)-O1P
Whole simulation	Mean	61.93	87.31	92.79	104.81	110.21	88.90	67.10	88.53	94.86	98.09	93.26	101.37
	SD	1.58	4.66	4.57	5.70	5.82	4.03	1.75	4.96	4.99	5.18	4.86	6.04
First ns	Mean	62.06	86.90	93.17	105.04	108.80	89.76	67.10	90.65	95.76	97.39	93.54	98.31
	SD	1.61	4.29	4.68	5.70	6.32	4.30	1.71	5.00	5.21	5.08	5.00	5.65
Last ns	Mean	61.88	86.63	92.42	104.07	107.25	89.40	67.22	89.49	96.88	97.96	94.30	100.18
	SD	1.52	4.54	4.46	5.60	5.53	3.99	1.77	5.16	4.78	5.10	5.04	5.81

Table 2. Mg coordination angles

Figure 7.13: **Mg ion coordination in the PheA-Tyr simulation.** The distances (table 1) and angles (table 2) between the magnesium ion and the six ligands in the distorted octahedral geometry; OE1 and OE2 from Glu 311, O1P and O2P from AMP, Thr 174 OG and Thr 310 OG atoms.



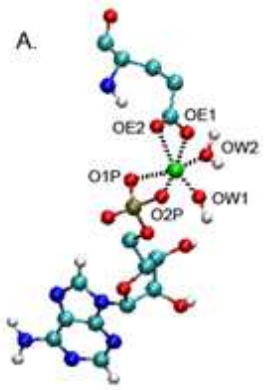
	Bond	Mg-OG1(1)	Mg-OG1(2)	Mg-OE1	Mg-OE2	Mg-O1P	Mg-O2P
Whole simulation	Mean	0.199	0.201	0.207	0.206	0.198	0.199
	SD	0.007	0.008	0.006	0.006	0.004	0.005
First ns	Mean	0.201	0.201	0.207	0.207	0.197	0.199
	SD	0.008	0.008	0.006	0.006	0.004	0.005
Last ns	Mean	0.198	0.202	0.206	0.207	0.198	0.199
	SD	0.007	0.008	0.006	0.006	0.004	0.005

Table 1. Mg coordination distances

	Angle	OE1-OE2	OG1(1)-OG1(2)	OE2-OG1(2)	OE1-O1P	OE2-O2P	OG1(2)-O2P	O2P-O1P	O1P-OE2	OG1(1)-OE1	OG1(2)-OE1	OG1(1)-O2P	OG1(1)-O1P
Whole simulation	Mean	62.17	87.31	92.25	100.16	103.57	94.11	67.29	90.78	97.62	97.97	97.00	96.45
	SD	1.89	5.10	5.11	6.73	8.18	5.64	2.07	6.05	6.09	5.99	6.33	6.28
First ns	Mean	62.08	91.16	91.15	103.33	113.90	92.11	67.11	87.60	91.65	97.07	92.29	100.55
	SD	1.64	5.21	4.43	5.64	6.27	4.67	1.83	5.24	4.58	5.27	5.20	6.59
Last ns	Mean	61.84	86.90	94.56	103.78	103.93	90.39	67.12	89.08	96.71	96.93	96.53	96.56
	SD	3.65	6.45	6.87	8.08	8.24	6.28	4.02	6.79	6.74	7.39	7.42	7.52

Table 2. Mg coordination angles

Figure 7.14: **Mg ion coordination in the PheA-Asp simulation.** The distances (table 1) and angles (table 2) between the magnesium ion and the six ligands in the distorted octahedral geometry; OE1 and OE2 from Glu 311, O1P and O2P from AMP, Thr 174 OG and Thr 310 OG atoms.



	Bond	Mg-OE1	Mg-OE2	Mg-O1P	Mg-O2P	Mg-OW1	Mg-OW2
Whole simulation	Mean	0.206	0.210	0.199	0.202	0.196	0.196
	SD	0.007	0.008	0.006	0.007	0.007	0.007
First ns	Mean	0.206	0.213	0.200	0.202	0.197	0.196
	SD	0.006	0.008	0.005	0.005	0.006	0.006
Last ns	Mean	0.204	0.209	0.198	0.201	0.195	0.195
	SD	0.016	0.017	0.015	0.016	0.016	0.015

Table 1. Mg coordination distances

	Angle	OE1-OE2	OE1-OW1	OE1-O2P	OE1-O1P	OE2-OW2	OW2-O2P	O2P-O1P	O1P-OE2	OW1-OE2	OW1-OW2	OW1-O2P	OW2-O1P
Whole simulation	Mean	62.02	102.90	100.66	93.75	97.42	100.04	67.01	90.14	91.06	90.91	94.96	94.63
	SD	1.65	6.48	7.15	5.97	6.17	5.81	1.75	5.40	4.61	4.19	4.81	4.88
First ns	Mean	61.55	106.21	110.10	92.70	91.97	98.22	66.95	93.07	89.28	90.53	93.41	97.04
	SD	1.79	7.45	8.00	7.00	5.06	6.24	1.78	6.20	4.76	4.43	4.48	5.69
Last ns	Mean	62.07	103.12	100.35	93.54	97.18	100.35	66.98	89.59	91.05	91.36	94.93	94.79
	SD	1.64	6.24	6.98	5.69	6.05	6.02	1.78	5.27	4.71	4.35	4.67	5.14

Table 2. Mg coordination angles

Figure 7.15: **Mg ion coordination in the PheA-Arg simulation.** The distances (table 1) and angles (table 2) between the magnesium ion and the six ligands in the distorted octahedral geometry; OE1 and OE2 from Glu 311, O1P and O2P from AMP, and two water oxygen atoms.

Appendix IIII

7.2 Modeller Parameter and Run Files

7.2.1 Script to check alignment

```
log.level(output=1, notes=1, warnings=1, errors=1, memory=0)
env = environ()

aln = alignment(env)
aln.append(file='CchH2-1AMU-1MD9.ali', align_codes='all')
aln.write(file='CchH2-1AMU-1MD9.pap', alignment_format='PAP')
aln.write(file='CchH2-1AMU-1MD9.fasta', alignment_format='FASTA')
aln.check()
```

7.2.2 Script to build models

```
from modeller.automodel import *

log.verbose()
env = environ()
env.io.hetatm = env.io.water = True
a = automodel(env, alnfile='CchH2-1AMU-1MD9.ali',
              knowns=('1AMU','1MD9'), sequence='CchH2')
a.starting_model = 1
a.ending_model = 100
a.make()
```

7.2.3 Script to assess model using ga341

```
score = mdl.assess_ga341()

from modeller.automodel import *    # Load the automodel class

log.verbose()    # request verbose output
env = environ()  # create a new MODELLER environment to build this model in
env.libs.topology.read(file='$(LIB)/top_heav.lib') # read topology
```

```
env.libs.parameters.read(file='$(LIB)/par.lib') # read parameters

# directories for input atom files
env.io.atom_files_directory = './:../atom_files'

# read model file
mdl = model(env)
mdl.read(file='CchH2.BEST80.pdb')
aln = alignment(env)
code = "CchH2"
# generate topology
aln.append_model(mdl, atom_files='CchH2.BEST80.pdb', align_codes=code)
aln.append_model(mdl, atom_files='CchH2.BEST80.pdb', align_codes=code+'-ini')
mdl.generate_topology(aln, sequence=code+'-ini')
mdl.transfer_xyz(aln)

mdl.assess_dope(output='ENERGY_PROFILE NO_REPORT', file='CchH2.BEST80.profile',
                normalize_profile=True, smoothing_window=15)
```

7.2.4 Add in and refine loop residues

```
#loop refinement of an existing model

from modeller.automodel import *

log.verbose()
env = environ()

# directories for input atom files
env.io.atom_files_directory = './:../atom_files'

# Create a new class based on 'loopmodel' so that we can redefine
# select_loop_atoms (necessary)
class myloop(loopmodel):
    # This routine picks the residues to be refined by loop modeling
    def select_loop_atoms(self):
        # 4 residue insertion (1st loop)
        self.pick_atoms(selection_segment=('175:', '181:'),
                        selection_status='INITIALIZE')

m = myloop(env,
```

```
            inimodel='model.pdb', # initial model of the target
            sequence='PheA')               # code of the target

m.loop.starting_model= 1           # index of the first loop model
m.loop.ending_model  = 100           # index of the last loop model
m.loop.md_level = refine.very_slow   # loop refinement method

m.make()
```

7.3 AMP Hydrogen Bonding

7.4 Magnesium Ion Coordination in the CchH2 holo simulations

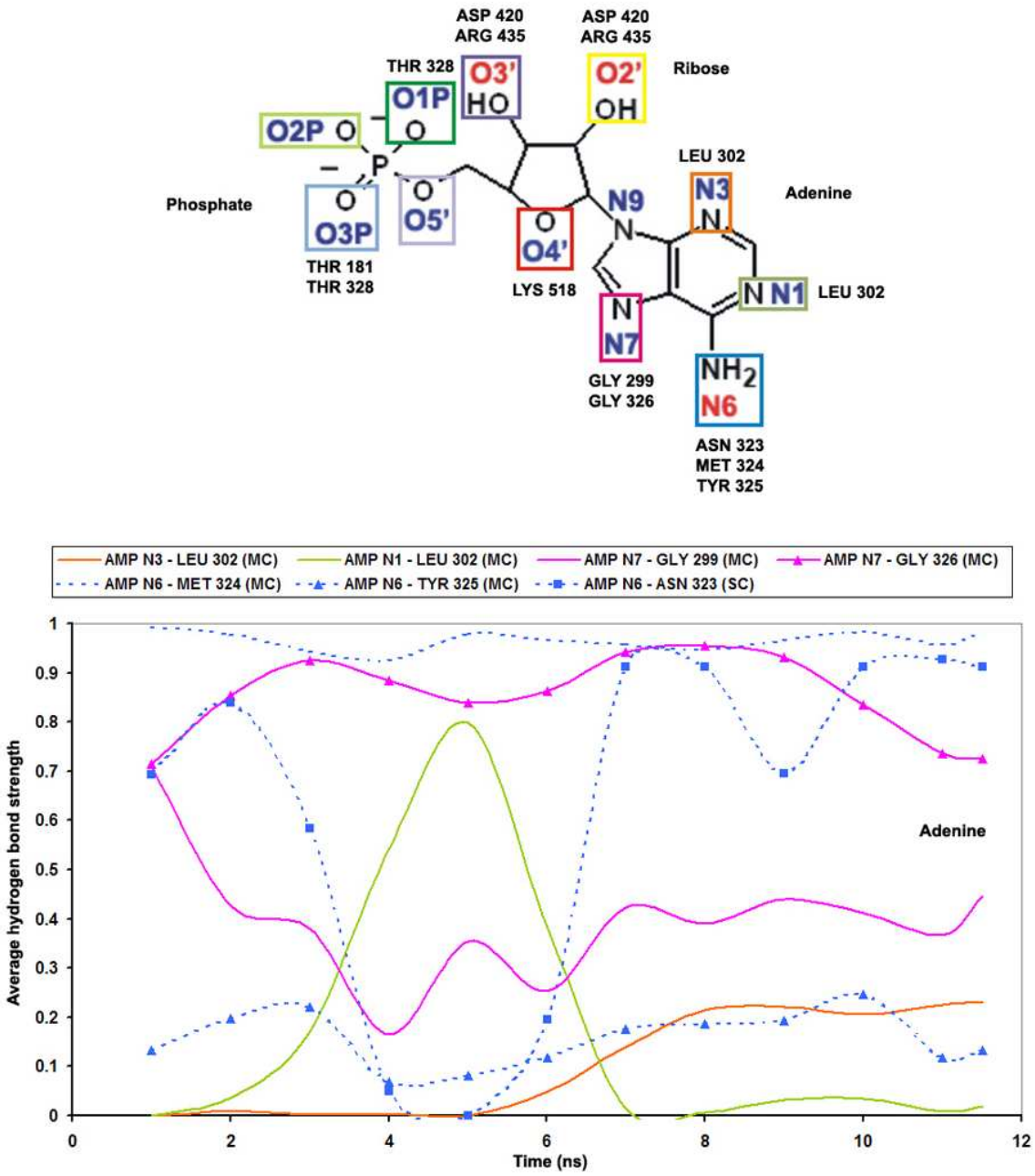


Figure 7.16: **Hydrogen bonding between AMP cofactor and CchH2; CchH2-Thr simulation.** Top: The structure of the AMP cofactor annotated with the CchH2 residues it forms hydrogen bonding interactions with. The AMP atom labels are coloured according to whether they are a potential acceptor (blue) or donor and acceptor (red) atom in the hydrogen bonding interaction. Each atom is framed by a different coloured box. This colour coding is used in the hydrogen bonding graphs to denote the hydrogen bonding interactions of each individual AMP atom. The hydrogen bonding interactions between each part of the AMP molecule; Adenine, Ribose and Phosphate, and the CchH2 protein have been considered separately. Bottom: Graph to show the hydrogen bonding interactions between the Adenine group of the AMP cofactor and the CchH2 protein as a function of time. Solid lines represent interactions where the AMP atom is an acceptor and broken lines where the AMP atom is acting as a donor. Hydrogen bonds are measured as the average strength per ns, calculated from data obtained every 1 ps, and are plotted at the ns marker.

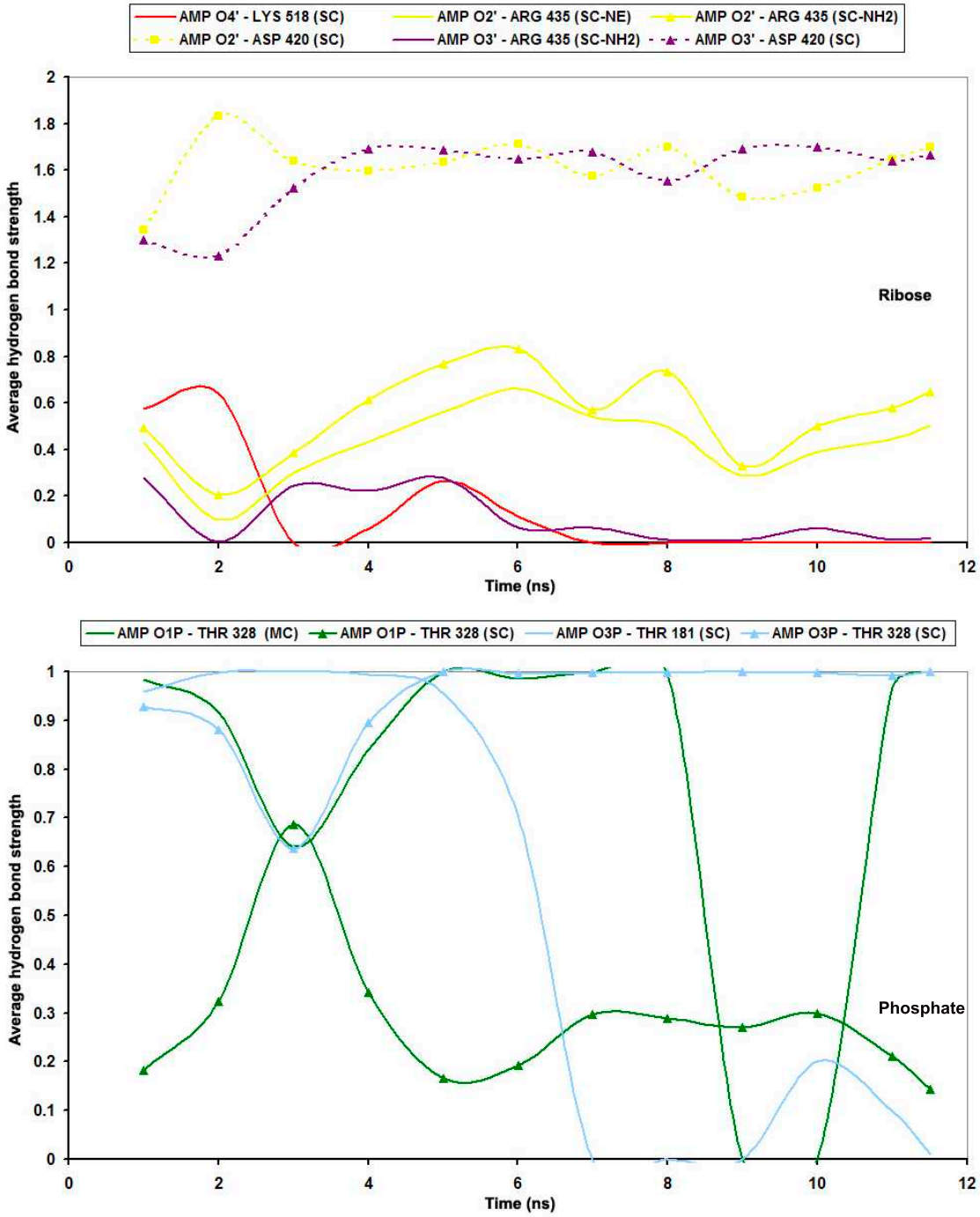


Figure 7.17: **Hydrogen bonding between AMP Ribose and Phosphate groups, and CchH2; CchH2-Thr simulation.** Top: Graphs to show the hydrogen bonding interactions between the Ribose (top graph) and the Phosphate (bottom graph) groups of the AMP cofactor and the CchH2 protein; as a function of time. Solid lines represent interactions where the AMP atom is an acceptor and broken lines where the AMP atom is acting as a donor. Hydrogen bonds are measured as the average strength per ns, calculated from data obtained every 1 ps, and are plotted at the ns marker.

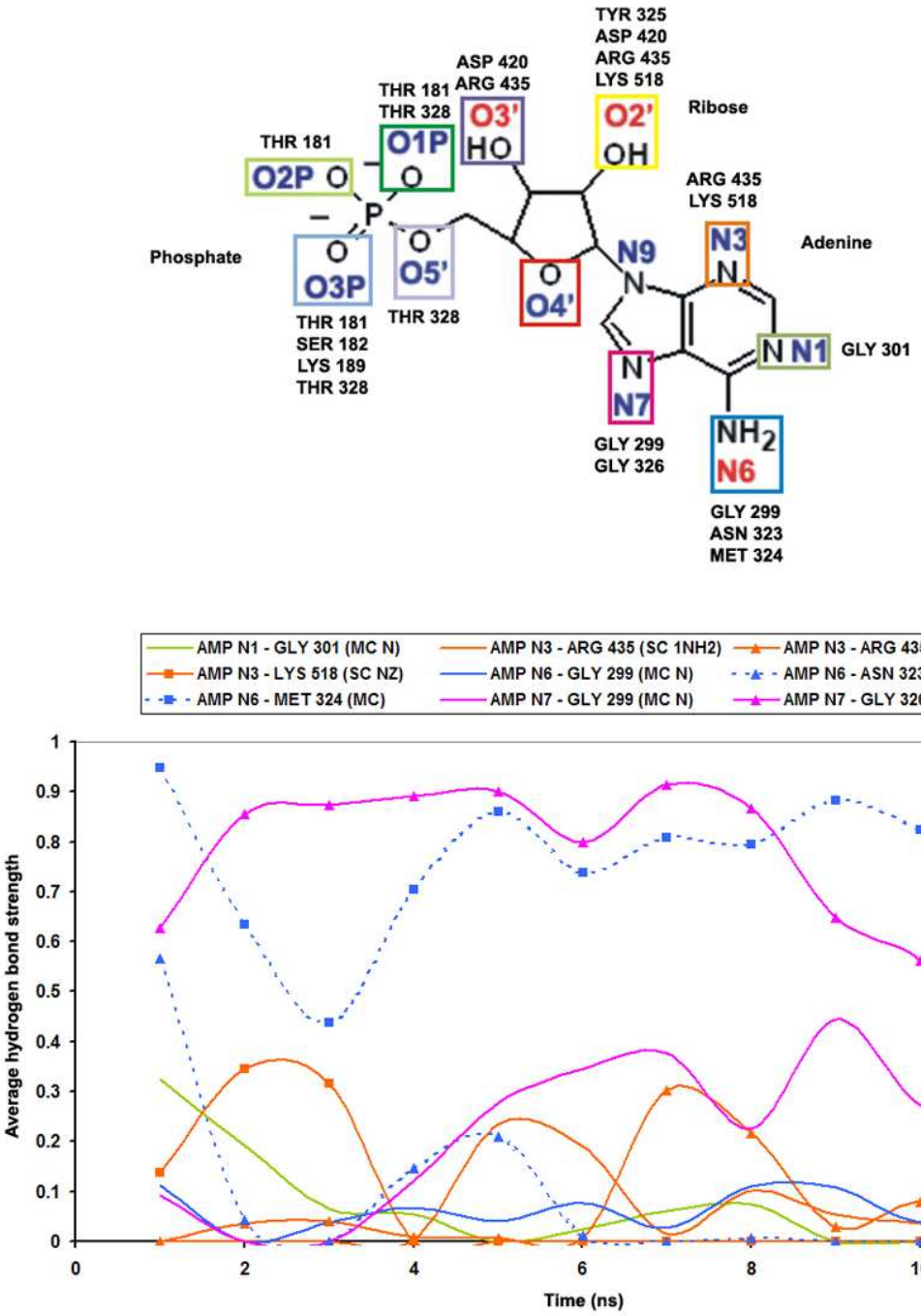


Figure 7.18: **Hydrogen bonding between AMP and CchH2; CchH2-Ser simulation.** Top: The structure of the AMP cofactor annotated with the CchH2 residues it forms hydrogen bonding interactions with. The AMP atom labels are coloured according to whether they are a potential acceptor (blue) or donor and acceptor (red) atom in the hydrogen bonding interaction. Each atom is framed by a different coloured box. This colour coding is used in the hydrogen bonding graphs to denote the hydrogen bonding interactions of each individual AMP atom. The hydrogen bonding interactions between each part of the AMP molecule; Adenine, Ribose and Phosphate, and the CchH2 protein have been considered separately. Bottom: Graph to show the hydrogen bonding interactions between the Adenine group of the AMP cofactor and the CchH2 protein as a function of time. Solid lines represent interactions where the AMP atom is an acceptor and broken lines where the AMP atom is acting as a donor. Hydrogen bonds are measured as the average strength per ns, calculated from data obtained every 1 ps, and are plotted at the ns marker.

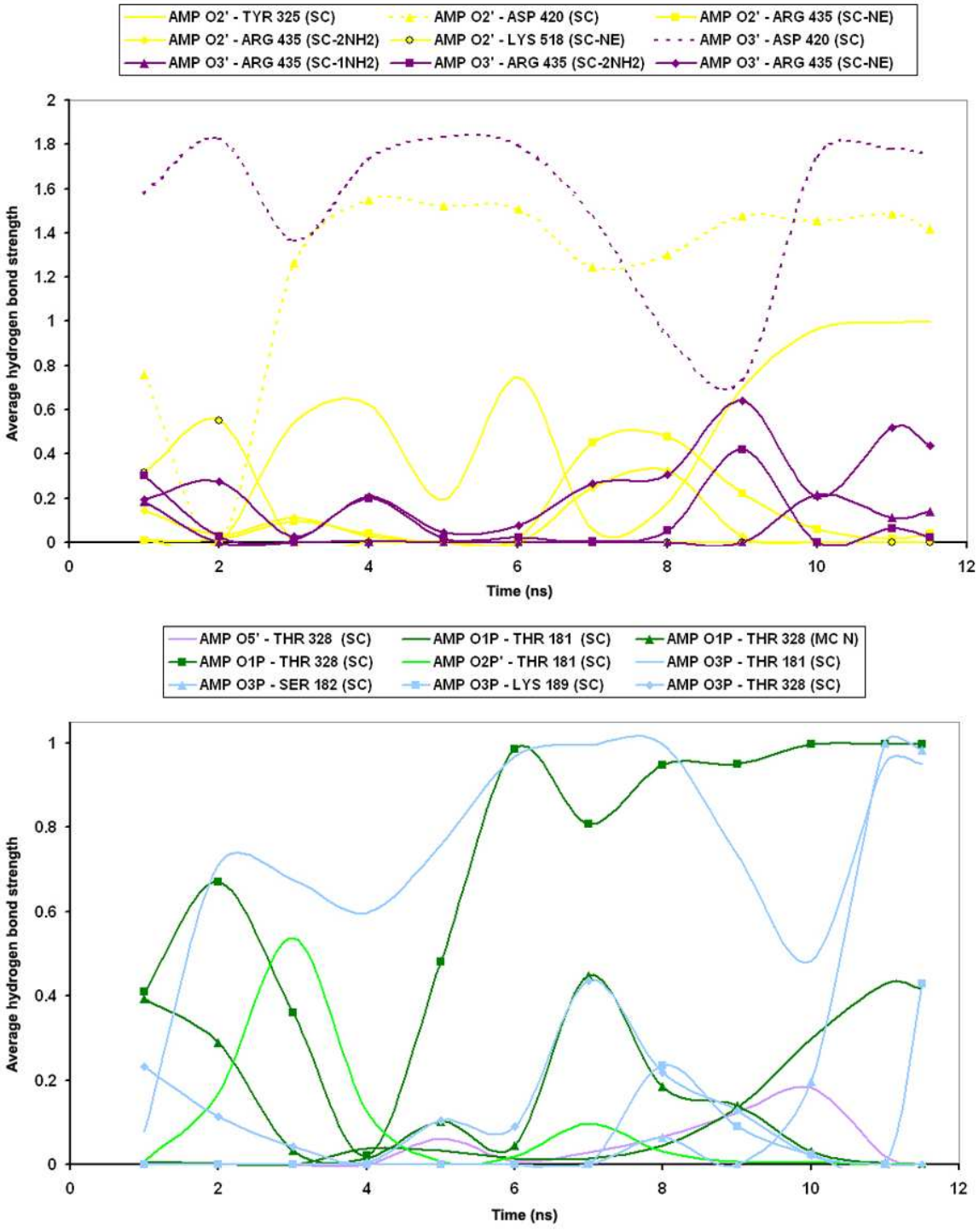


Figure 7.19: **Hydrogen bonding between AMP Ribose and Phosphate groups, and CchH2; CchH2-Ser simulation.** Top: Graphs to show the hydrogen bonding interactions between the Ribose (top graph) and the Phosphate (bottom graph) groups of the AMP cofactor and the CchH2 protein; as a function of time. Solid lines represent interactions where the AMP atom is an acceptor and broken lines where the AMP atom is acting as a donor. Hydrogen bonds are measured as the average strength per ns, calculated from data obtained every 1 ps, and are plotted at the ns marker.

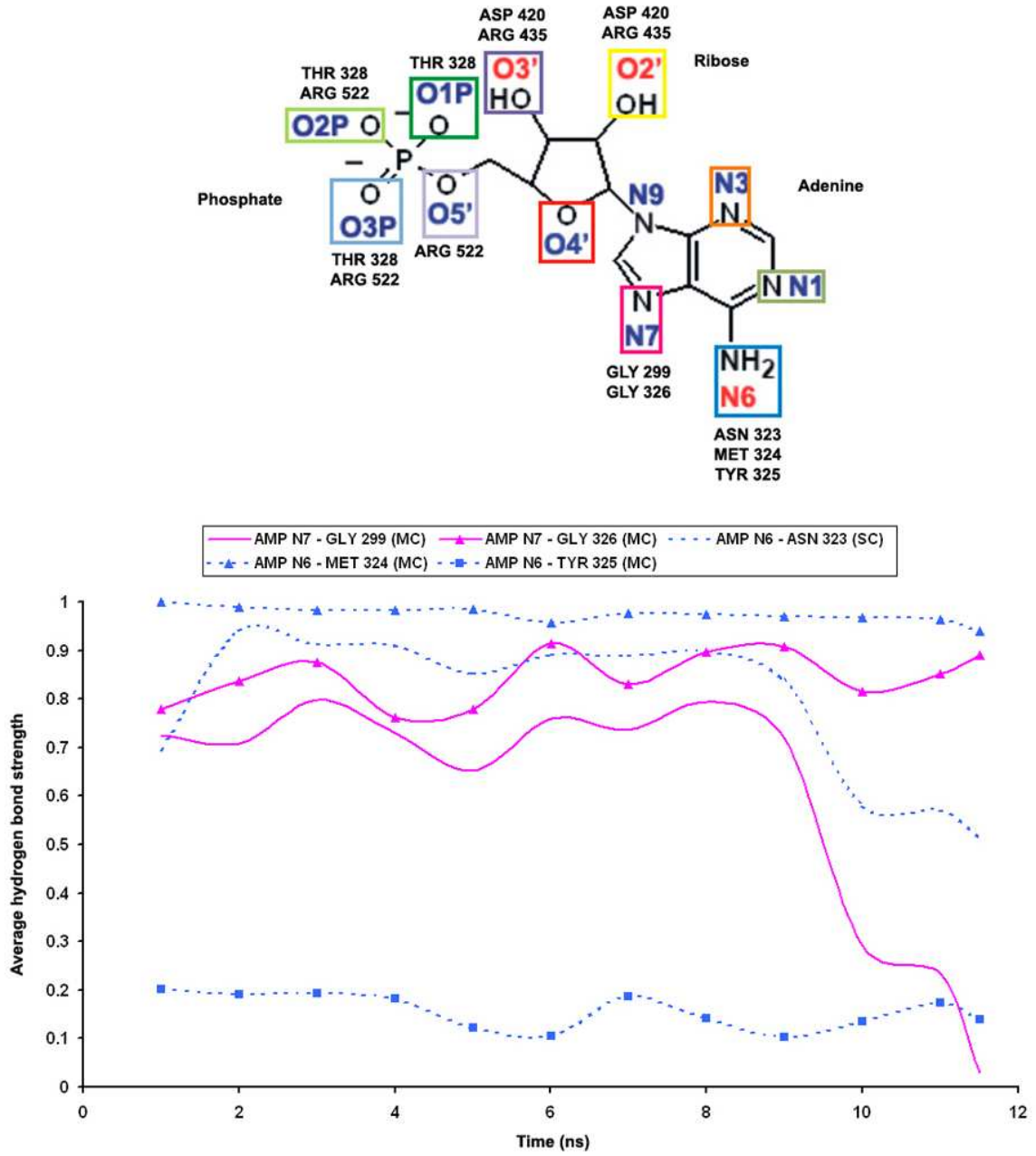


Figure 7.20: **Hydrogen bonding between AMP and CchH2; CchH2-Val simulation.** Top: The structure of the AMP cofactor annotated with the CchH2 residues it forms hydrogen bonding interactions with. The AMP atom labels are coloured according to whether they are a potential acceptor (blue) or donor and acceptor (red) atom in the hydrogen bonding interaction. Each atom is framed by a different coloured box. This colour coding is used in the hydrogen bonding graphs to denote the hydrogen bonding interactions of each individual AMP atom. The hydrogen bonding interactions between each part of the AMP molecule; Adenine, Ribose and Phosphate, and the CchH2 protein have been considered separately. Bottom: Graph to show the hydrogen bonding interactions between the Adenine group of the AMP cofactor and the CchH2 protein as a function of time. Solid lines represent interactions where the AMP atom is an acceptor and broken lines where the AMP atom is acting as a donor. Hydrogen bonds are measured as the average strength per ns, calculated from data obtained every 1 ps, and are plotted at the ns marker.

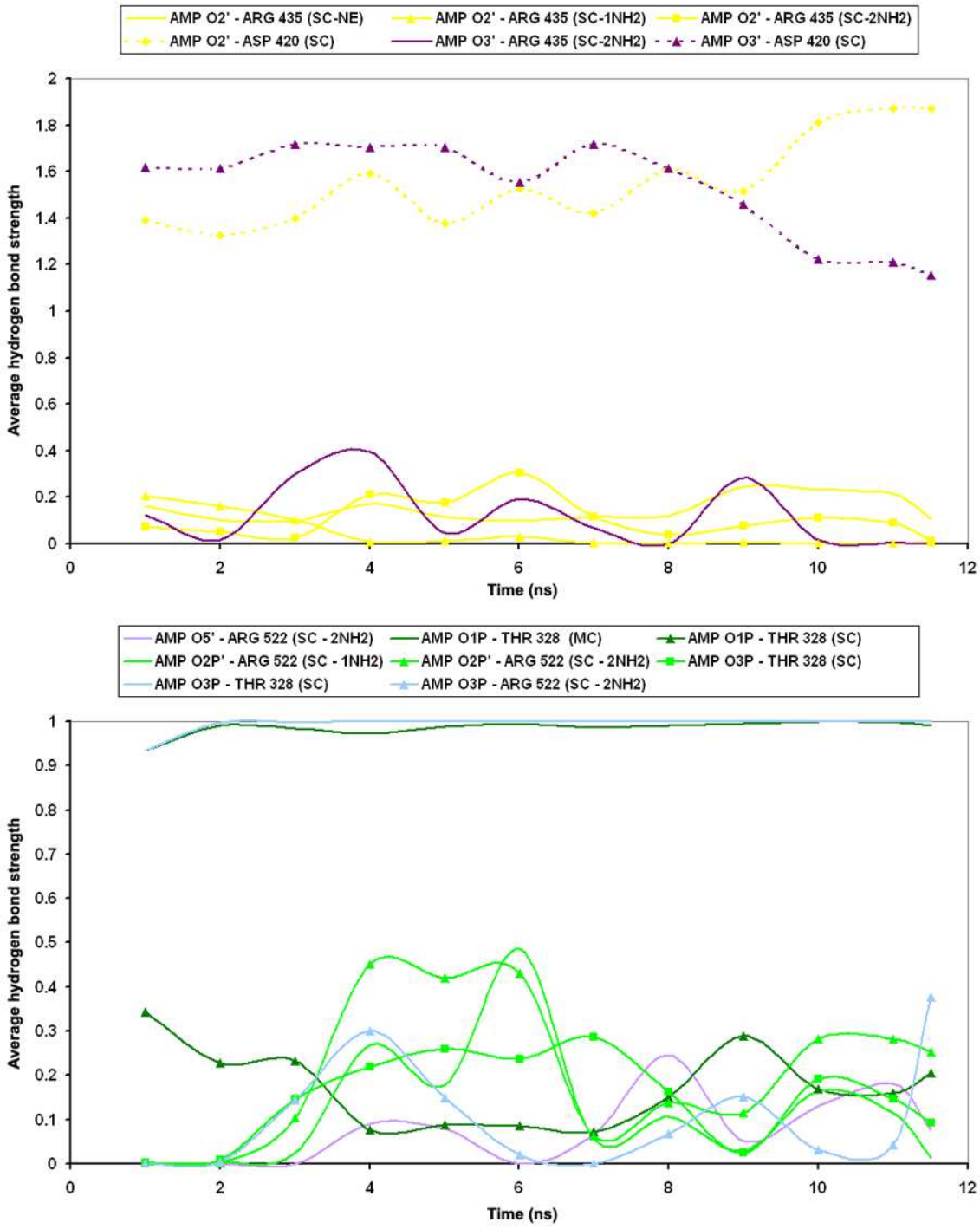


Figure 7.21: **Hydrogen bonding between AMP Ribose and Phosphate groups, and CchH2; CchH2-Val simulation.** Top: Graphs to show the hydrogen bonding interactions between the Ribose (top graph) and the Phosphate (bottom graph) groups of the AMP cofactor and the CchH2 protein; as a function of time. Solid lines represent interactions where the AMP atom is an acceptor and broken lines where the AMP atom is acting as a donor. Hydrogen bonds are measured as the average strength per ns, calculated from data obtained every 1 ps, and are plotted at the ns marker.

	Bond	Mg-OE1	Mg-OE2	Mg-O1P	Mg-O2P	Mg-OW1	Mg-OW2
Whole simulation	Mean	0.208	0.208	0.200	0.200	0.197	0.196
	SD	0.006	0.006	0.005	0.005	0.006	0.006
First ns	Mean	0.209	0.206	0.201	0.201	0.196	0.196
	SD	0.006	0.006	0.005	0.005	0.006	0.006
Last ns	Mean	0.209	0.208	0.200	0.200	0.196	0.196
	SD	0.007	0.006	0.005	0.005	0.005	0.005

Table 1. Mg coordination distances

	Angle	OE1-OE2	OE1-OW2	OE1-O2P	OE1-O1P	OE2-OW2	OW2-O2P	O2P-O1P	O1P-OE2	OW1-OE2	OW1-OW2	OW1-O2P	OW1-O1P
Whole simulation	Mean	62.01	92.77	98.66	89.31	96.35	95.05	67.32	100.12	97.14	90.88	101.87	93.64
	SD	1.64	5.01	6.39	5.14	5.81	4.73	1.75	6.70	5.61	4.15	5.73	4.38
First ns	Mean	62.01	92.05	99.90	90.50	98.93	95.54	67.21	97.45	96.40	90.72	101.58	93.55
	SD	1.60	5.06	7.05	5.75	6.41	5.27	1.75	7.00	6.00	4.20	5.68	4.75
Last ns	Mean	62.24	92.74	96.83	88.33	95.84	95.25	67.42	99.96	98.07	91.64	102.34	93.44
	SD	1.72	4.93	5.64	4.87	5.32	4.78	1.70	6.51	5.67	4.12	5.76	4.15

Table 2. Mg coordination angles

Figure 7.22: **Mg ion coordination in the CchH2-Thr system simulation.** The distances (table 1) and angles (table 2) between the magnesium ion and the six ligands in the distorted octahedral geometry; OE1 and OE2 from Glu 329, O1P and O2P from AMP and two water oxygen atoms.

A.

Table 1. Mg coordination distances

	Bond	Mg-OE1	Mg-OE2	Mg-O1P	Mg-O2P	Mg-OW1	Mg-OW2
Whole simulation	Mean	0.209	0.206	0.201	0.201	0.196	0.197
	SD	0.006	0.006	0.005	0.005	0.006	0.006
First ns	Mean	0.209	0.206	0.202	0.200	0.197	0.196
	SD	0.006	0.006	0.006	0.005	0.006	0.006
Last ns	Mean	0.208	0.206	0.202	0.200	0.195	0.197
	SD	0.006	0.006	0.005	0.005	0.006	0.006

Table 2. Mg coordination angles

	Angle	OE1-OE2	OE1-OW2	OE1-O2P	OE1-O1P	OE2-OW2	OW2-O2P	O2P-O1P	O1P-OE2	OW1-OE2	OW1-OW2	OW1-O2P	OW1-O1P
Whole simulation	Mean	62.02	90.96	99.14	92.18	100.81	95.75	67.01	95.77	96.51	90.38	102.49	93.15
	SD	1.62	4.49	6.49	5.51	5.98	5.01	1.77	6.30	5.36	4.04	6.20	4.60
First ns	Mean	61.94	91.94	97.02	92.41	100.34	95.77	66.86	96.98	96.12	91.01	104.62	91.59
	SD	1.65	4.85	6.89	5.57	5.88	5.63	1.76	7.40	5.77	4.39	7.19	4.80
Last ns	Mean	62.25	90.75	97.40	94.05	100.02	96.93	66.98	96.09	95.45	90.27	105.04	91.40
	SD	1.55	4.35	5.60	5.48	5.89	5.12	1.70	5.87	4.56	3.80	5.67	3.87

Figure 7.23: **Mg ion coordination in the CchH2-Ser system simulation.** The distances (table 1) and angles (table 2) between the magnesium ion and the six ligands in the distorted octahedral geometry; OE1 and OE2 from Glu 329, O1P and O2P from AMP and two water oxygen atoms.

	Bond	Mg-OE1	Mg-OE2	Mg-O1P	Mg-O2P	Mg-OW1	Mg-OW2
Whole simulation	Mean	0.209	0.207	0.201	0.201	0.196	0.196
	SD	0.007	0.006	0.005	0.005	0.006	0.006
First ns	Mean	0.209	0.207	0.201	0.201	0.196	0.197
	SD	0.007	0.006	0.005	0.005	0.006	0.006
Last ns	Mean	0.210	0.208	0.201	0.201	0.196	0.197
	SD	0.007	0.006	0.005	0.005	0.006	0.006

Table 1. Mg coordination distances

	Angle	OE1-OE2	OE1-OW2	OE1-O2P	OE1-O1P	OE2-OW2	OW2-O2P	O2P-O1P	O1P-OE2	OW1-OE2	OW1-OW2	OW1-O2P	OW1-O1P
Whole simulation	Mean	62.03	93.69	98.88	87.51	95.80	94.20	67.01	101.55	96.98	91.32	101.78	94.74
	SD	1.64	5.71	6.77	5.22	6.02	4.70	1.76	7.30	5.69	4.31	6.67	4.94
First ns	Mean	61.99	92.76	101.03	89.35	97.04	94.03	67.10	100.66	96.03	91.43	100.76	93.99
	SD	1.63	5.21	6.93	5.44	6.67	4.79	1.76	7.76	5.61	4.43	5.60	4.91
Last ns	Mean	61.91	93.67	96.61	88.18	94.96	94.10	66.93	102.52	97.40	90.61	103.77	94.54
	SD	1.61	5.23	5.38	5.29	5.35	4.54	1.71	6.43	5.42	4.19	5.92	4.27

Table 2. Mg coordination angles

Figure 7.24: **Mg ion coordination in the CchH2-Val system simulation.** The distances (table 1) and angles (table 2) between the magnesium ion and the six ligands in the distorted octahedral geometry; OE1 and OE2 from Glu 329, O1P and O2P from AMP and two water oxygen atoms.

References

[1] Luo, L., Burkart, M., Stachehaus, T. & Walsh, C. T. Substrate recognition and selection by the initiation module pheate of gramicidin s synthetase. *Journal of the American Chemical Society* **123**, 11208–11218 (2001).

[2] von Dohren, H., Keller, U., Vater, J. & Zocher, R. Multifunctional peptide synthetases. *Chemical Reviews* **97**, 2675–2706 (1997).

[3] Konz, D. & Marahiel, M. A. How do peptide synthetases generate structural diversity? *Chem Biol* **6**, R39–48 (1999).

[4] Weissman, K. J. & Muller, R. Protein-protein interactions in multienzyme megasynthetases. *ChemBioChem* **9**, 826–848 (2008).

[5] Mootz, H. D., Schwarzer, D. & Marahiel, M. A. Ways of assembling complex natural products on modular nonribosomal peptide synthetases. *Chembiochem* **3**, 490–504 (2002).

[6] Schwarzer, D., Finking, R. & Marahiel, M. A. Nonribosomal peptides: from genes to products. *Nat Prod Rep* **20**, 275–87 (2003).

[7] Kleinkauf, H. & Von Dohren, H. A nonribosomal system of peptide biosynthesis. *Eur J Biochem* **236**, 335–51 (1996).

[8] Hori, K. *et al.* Molecular cloning and nucleotide sequence of the gramicidin s synthetase 1 gene. *J Biochem (Tokyo)* **106**, 639–45 (1989).

[9] Turgay, K., Krause, M. & Marahiel, M. A. Four homologous domains in the pri-

mary structure of grsb are related to domains in a superfamily of adenylate-forming enzymes. *Mol Microbiol* **6**, 2743–4 (1992).

[10] Saito, F., Hori, K., Kanda, M., Kurotsu, T. & Saito, Y. Entire nucleotide sequence for bacillus brevis nagano grs2 gene encoding gramicidin s synthetase 2: a multifunctional peptide synthetase. *J Biochem* **116**, 357–67 (1994).

[11] van Wageningen, A. M. *et al.* Sequencing and analysis of genes involved in the biosynthesis of a vancomycin group antibiotic. *Chem Biol* **5**, 155–62 (1998).

[12] Eliopoulos, G. M. *et al.* In vitro and in vivo activity of ly 146032, a new cyclic lipopeptide antibiotic. *Antimicrob Agents Chemother* **30**, 532–5 (1986).

[13] Tally, F. P. *et al.* Daptomycin: a novel agent for gram-positive infections. *Expert Opin Investig Drugs* **8**, 1223–38 (1999).

[14] Tally, F. P. & DeBruin, M. F. Development of daptomycin for gram-positive infections. *J Antimicrob Chemother* **46**, 523–6 (2000).

[15] Byford, M. F., Baldwin, J. E., Shiau, C. Y. & Schofield, C. J. The mechanism of acv synthetase. *Chemistry Reviews* **97**, 2631–2650 (1997).

[16] Weber, G., Schorgendorfer, K., Schneider-Scherzer, E. & Leitner, E. The peptide synthetase catalyzing cyclosporine production in tolypocladium niveum is encoded by a giant 45.8-kilobase open reading frame. *Curr Genet* **26**, 120–5 (1994).

[17] Romero, F. *et al.* Thiocoraline, a new depsipeptide with antitumor activity produced by a marine micromonospora. i. taxonomy, fermentation, isolation, and biological activities. *J Antibiot (Tokyo)* **50**, 734–7 (1997).

[18] Singh, R., Sharma, M., Joshi, P. & Rawat, D. S. Clinical status of anti-cancer agents derived from marine sources. *Anticancer Agents Med Chem* **8**, 603–17 (2008).

[19] Gehring, A. M. *et al.* Iron acquisition in plague: modular logic in enzymatic biogenesis of yersiniabactin by yersinia pestis. *Chem Biol* **5**, 573–86 (1998).

[20] Keating, T. A., Marshall, C. G. & Walsh, C. T. Reconstitution and characterization of the vibrio cholerae vibriobactin synthetase from vibb, vibe, vibf, and vibh. *Biochemistry* **39**, 15522–30 (2000).

[21] Gehring, A. M., Mori, I. & Walsh, C. T. Reconstitution and characterization of the escherichia coli enterobactin synthetase from entb, ente, and entf. *Biochemistry* **37**, 2648–59 (1998).

[22] Marahiel, M. A., Stachelhaus, T. & Mootz, H. D. Modular peptide synthetases involved in nonribosomal peptide synthesis. *Chem Rev* **97**, 2651–2674 (1997). 0009-2665 (Print) Journal article.

[23] Lautru, S., Deeth, R. J., Bailey, L. M. & Challis, G. L. Discovery of a new peptide natural product by streptomyces coelicolor genome mining. *Nat Chem Biol* **1**, 265–9 (2005).

[24] Mootz, H. D. & Marahiel, M. A. The tyrocidine biosynthesis operon of bacillus brevis: complete nucleotide sequence and biochemical characterization of functional internal adenylation domains. *J Bacteriol* **179**, 6843–50 (1997).

[25] Rouhiainen, L. *et al.* Genes encoding synthetases of cyclic depsipeptides, anabaenopeptilides, in anabaena strain 90. *Mol Microbiol* **37**, 156–67 (2000).

[26] Lipmann, F. Bacterial production of antibiotic polypeptides by thiol-linked synthesis on protein templates. *Adv Microb Physiol* **21**, 227–66 (1980).

[27] Stein, T. *et al.* Detection of 4'-phosphopantetheine at the thioester binding site for l-valine of gramicidins synthetase 2. *FEBS Lett* **340**, 39–44 (1994).

[28] Stein, T. *et al.* The multiple carrier model of nonribosomal peptide biosynthesis at modular multienzymatic templates. *J Biol Chem* **271**, 15428–35 (1996).

[29] Marshall, C. G., Burkart, M. D., Meray, R. K. & Walsh, C. T. Carrier protein recognition in siderophore-producing nonribosomal peptide synthetases. *Biochemistry* **41**, 8429–37 (2002).

[30] Lautru, S. & Challis, G. L. Substrate recognition by nonribosomal peptide synthetase multi-enzymes. *Microbiology* **150**, 1629–36 (2004).

[31] Belshaw, P. J., Walsh, C. T. & Stachelhaus, T. Aminoacyl-coas as probes of condensation domain selectivity in nonribosomal peptide synthesis. *Science* **284**, 486–9 (1999).

[32] Miller, D. A., Luo, L., Hillson, N., Keating, T. A. & Walsh, C. T. Yersiniabactin synthetase: a four-protein assembly line producing the nonribosomal peptide/polyketide hybrid siderophore of yersinia pestis. *Chem Biol* **9**, 333–44 (2002).

[33] Trauger, J. W., Kohli, R. M., Mootz, H. D., Marahiel, M. A. & Walsh, C. T. Peptide cyclization catalysed by the thioesterase domain of tyrocidine synthetase. *Nature* **407**, 215–8 (2000).

[34] Linne, U. & Marahiel, M. A. Control of directionality in nonribosomal peptide synthesis: role of the condensation domain in preventing misinitiation and timing of epimerization. *Biochemistry* **39**, 10439–47 (2000).

[35] Ehmann, D. E., Shaw-Reid, C. A., Losey, H. C. & Walsh, C. T. The entf and ente adenylation domains of escherichia coli enterobactin synthetase: sequestration and selectivity in acyl-amp transfers to thiolation domain cosubstrates. *Proc Natl Acad Sci U S A* **97**, 2509–14 (2000).

[36] Doekel, S. & Marahiel, M. A. Dipeptide formation on engineered hybrid peptide synthetases. *Chem Biol* **7**, 373–84 (2000).

[37] Samel, S. A., Schoenafinger, G., Knappe, T. A., Marahiel, M. A. & Essen, L. O. Structural and functional insights into a peptide bond-forming bidomain from a nonribosomal peptide synthetase. *Structure* **15**, 781–92 (2007).

[38] Tanovic, A., Samel, S. A., Essen, L. O. & Marahiel, M. A. Crystal structure of the termination module of a nonribosomal peptide synthetase. *Science* **321**, 659–63 (2008).

[39] Cane, D. E. & Walsh, C. T. The parallel and convergent universes of polyketide synthases and nonribosomal peptide synthetases. *Chem Biol* **6**, R319–25 (1999).

[40] Du, L., Sanchez, C., Chen, M., Edwards, D. J. & Shen, B. The biosynthetic gene cluster for the antitumor drug bleomycin from streptomyces verticillus atcc15003 supporting functional interactions between nonribosomal peptide synthetases and a polyketide synthase. *Chem Biol* **7**, 623–42 (2000).

[41] Molnar, I. *et al.* The biosynthetic gene cluster for the microtubule-stabilizing agents epothilones a and b from sorangium cellulosum so ce90. *Chem Biol* **7**, 97–109 (2000).

[42] Tang, L. *et al.* Cloning and heterologous expression of the epothilone gene cluster. *Science* **287**, 640–2 (2000).

[43] Challis, G. L. & Naismith, J. H. Structural aspects of non-ribosomal peptide biosynthesis. *Curr Opin Struct Biol* **14**, 748–56 (2004).

[44] Taubes, G. The bacteria fight back. *Science* **321**, 356–361 (2008).

[45] Stachelhaus, T., Schneider, A. & Marahiel, M. A. Rational design of peptide antibiotics by targeted replacement of bacterial and fungal domains. *Science* **269**, 69–72 (1995).

[46] Mootz, H. D., Schwarzer, D. & Marahiel, M. A. Construction of hybrid peptide synthetases by module and domain fusions. *Proc Natl Acad Sci U S A* **97**, 5848–53 (2000).

[47] Mootz, H. D. *et al.* Decreasing the ring size of a cyclic nonribosomal peptide antibiotic by in-frame module deletion in the biosynthetic genes. *J Am Chem Soc* **124**, 10980–1 (2002).

[48] Eppelmann, K., Stachelhaus, T. & Marahiel, M. A. Exploitation of the selectivity-conferring code of nonribosomal peptide synthetases for the rational design of novel peptide antibiotics. *Biochemistry* **41**, 9718–26 (2002).

[49] Hahn, M. & Stachelhaus, T. Harnessing the potential of communication-mediating domains for the biocombinatorial synthesis of nonribosomal peptides. *Proc Natl Acad Sci U S A* **103**, 275–80 (2006).

[50] Babbitt, P. C. *et al.* Ancestry of the 4-chlorobenzoate dehalogenase: analysis of amino acid sequence identities among families of acyl:adenyl ligases, enoyl-coa hydratases/isomerases, and acyl-coa thioesterases. *Biochemistry* **31**, 5594–604 (1992).

[51] Scholten, J. D. *et al.* Novel enzymic hydrolytic dehalogenation of a chlorinated aromatic. *Science* **253**, 182–5 (1991).

[52] Gulick, A. M., Starai, V. J., Horswill, A. R., Homick, K. M. & Escalante-Semerena, J. C. The 1.75 a crystal structure of acetyl-coa synthetase bound to adenosine-5'-propylphosphate and coenzyme a. *Biochemistry* **42**, 2866–73 (2003).

[53] Reger, A. S., Wu, R., Dunaway-Mariano, D. & Gulick, A. M. Structural characterization of a 140 degrees domain movement in the two-step reaction catalyzed by 4-chlorobenzoate:coa ligase. *Biochemistry* **47**, 8016–25 (2008).

[54] Branchini, B. R. *et al.* Mutagenesis evidence that the partial reactions of firefly bioluminescence are catalyzed by different conformations of the luciferase c-terminal domain. *Biochemistry* **44**, 1385–93 (2005).

[55] van Liempt, H., Pfeifer, E., Schwecke, T., Palissa, H. & von Doehren, H. Principles of the molecular construction of multienzyme templates for peptide biosynthesis in integrated reaction sequences. *Biomed Biochim Acta* **50**, S256–9 (1991).

[56] Babbitt, P. C. Procite release 7. Tech. Rep. (1991).

[57] Walker, J. E., Saraste, M., Runswick, M. J. & Gay, N. J. Distantly related sequences in the alpha- and beta-subunits of atp synthase, myosin, kinases and other atp-requiring enzymes and a common nucleotide binding fold. *Embo J* **1**, 945–51 (1982).

[58] Fry, D. C., Kuby, S. A. & Mildvan, A. S. Nmr studies of the mgatp binding site of

adenylate kinase and of a 45-residue peptide fragment of the enzyme. *Biochemistry* **24**, 4680–94 (1985).

[59] Saraste, M., Sibbald, P. R. & Wittinghofer, A. The p-loop–a common motif in atp- and gtp-binding proteins. *Trends Biochem Sci* **15**, 430–4 (1990).

[60] Smith, C. A. & Rayment, I. Active site comparisons highlight structural similarities between myosin and other p-loop proteins. *Biophys J* **70**, 1590–602 (1996).

[61] Conti, E., Franks, N. P. & Brick, P. Crystal structure of firefly luciferase throws light on a superfamily of adenylate-forming enzymes. *Structure* **4**, 287–98 (1996).

[62] Conti, E., Stachelhaus, T., Marahiel, M. A. & Brick, P. Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin s. *Embo J* **16**, 4174–83 (1997).

[63] Gulick, A. M., Lu, X. & Dunaway-Mariano, D. Crystal structure of 4-chlorobenzoate:coa ligase/synthetase in the unliganded and aryl substrate-bound states. *Biochemistry* **43**, 8670–9 (2004).

[64] von Dohren, H. The organization of multifunctional peptide and depsipeptide synthetases. *Biochem Soc Trans* **21**, 214–7 (1993).

[65] Rapaport, E., Remy, P., Kleinkauf, H., Vater, J. & Zamecnik, P. C. Aminoacyl-trna synthetases catalyze amp—-adp—-atp exchange reactions, indicating labile covalent enzyme-amino-acid intermediates. *Proc Natl Acad Sci U S A* **84**, 7891–5 (1987).

[66] Eriani, G., Dirheimer, G. & Gangloff, J. Aspartyl-trna synthetase from escherichia coli: cloning and characterisation of the gene, homologies of its translated amino acid sequence with asparaginyl- and lysyl-trna synthetases. *Nucleic Acids Res* **18**, 7109–18 (1990).

[67] Peypoux, F. *et al.* Isolation and characterization of a new variant of surfactin, the [val7]surfactin. *Eur J Biochem* **202**, 101–6 (1991).

[68] Peypoux, F. *et al.* [ala4]surfactin, a novel isoform from bacillus subtilis studied by mass and nmr spectroscopies. *Eur J Biochem* **224**, 89–96 (1994).

[69] Lawen, A. & Traber, R. Substrate specificities of cyclosporin synthetase and peptolide sdz 214-103 synthetase. comparison of the substrate specificities of the related multifunctional polypeptides. *J Biol Chem* **268**, 20452–65 (1993).

[70] Kleinkauf, H., Dittmann, J. & Lawen, A. Cell-free biosynthesis of cyclosporin a and analogues. *Biomed Biochim Acta* **50**, S219–24 (1991).

[71] Pieper, R., Kleinkauf, H. & Zocher, R. Enniatin synthetases from different fusaria exhibiting distinct amino acid specificities. *J Antibiot (Tokyo)* **45**, 1273–7 (1992).

[72] Galli, G. *et al.* Characterization of the surfactin synthetase multi-enzyme complex. *Biochim Biophys Acta* **1205**, 19–28 (1994).

[73] May, J. J., Kessler, N., Marahiel, M. A. & Stubbs, M. T. Crystal structure of dhbe, an archetype for aryl acid activating domains of modular nonribosomal peptide synthetases. *Proc Natl Acad Sci U S A* **99**, 12120–5 (2002).

[74] Dieckmann, R., Lee, Y. O., van Liempt, H., von Dohren, H. & Kleinkauf, H. Expression of an active adenylate-forming domain of peptide synthetases corresponding to acyl-coa-synthetases. *FEBS Letters* **357**, 212–6 (1995).

[75] Dieckmann, R., Pavela-Vrancic, M., von Dohren, H. & Kleinkauf, H. Probing the domain structure and ligand-induced conformational changes by limited proteolysis of tyrocidine synthetase 1. *J Mol Biol* **288**, 129–40 (1999).

[76] Stachelhaus, T. & Marahiel, M. A. Modular structure of peptide synthetases revealed by dissection of the multifunctional enzyme grsa. *J Biol Chem* **270**, 6163–9 (1995).

[77] Crosa, J. H. & Walsh, C. T. Genetics and assembly line enzymology of siderophore biosynthesis in bacteria. *Microbiology Molecular Biology Review* **66**, 223–49 (2002).

[78] Kessler, N., Schuhmann, H., Morneweg, S., Linne, U. & Marahiel, M. A. The linear pentadecapeptide gramicidin is assembled by four multimodular nonribosomal peptide synthetases that comprise 16 modules with 56 catalytic domains. *J Biol Chem* **279**, 7413–9 (2004).

[79] Pavela-Vrancic, M., Pfeifer, E., Schroder, W., von Dohren, H. & Kleinkauf, H. Identification of the atp binding site in tyrocidine synthetase 1 by selective modification with fluorescein 5'-isothiocyanate. *J Biol Chem* **269**, 14962–6 (1994).

[80] Hamoen, L. W., Eshuis, H., Jongbloed, J., Venema, G. & van Sinderen, D. A small gene, designated coms, located within the coding region of the fourth amino acid-activation domain of srfa, is required for competence development in bacillus subtilis. *Mol Microbiol* **15**, 55–63 (1995).

[81] Stachelhaus, T., Mootz, H. D. & Marahiel, M. A. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem Biol* **6**, 493–505 (1999).

[82] Challis, G. L., Ravel, J. & Townsend, C. A. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chemistry and Biology* **7**, 211–24 (2000).

[83] Cerdeno, A. M., Bibb, M. J. & Challis, G. L. Analysis of the prodiginine biosynthesis gene cluster of streptomyces coelicolor a3(2): new mechanisms for chain initiation and termination in modular multienzymes. *Chemistry and Biology* **8**, 817–29 (2001).

[84] Thomas, M. G., Burkart, M. D. & Walsh, C. T. Conversion of l-proline to pyrrolyl-2-carboxyl-s-pcp during undecylprodigiosin and pyoluteorin biosynthesis. *Chem Biol* **9**, 171–84 (2002).

[85] Ackerley, D. F., Caradoc-Davies, T. T. & Lamont, I. L. Substrate specificity of the nonribosomal peptide synthetase pvdd from pseudomonas aeruginosa. *Journal of Bacteriology* **185**, 2848–55 (2003).

[86] Franks, N. P., Jenkins, A., Conti, E., Lieb, W. R. & Brick, P. Structural basis for the inhibition of firefly luciferase by a general anesthetic. *Biophys J* **75**, 2205–11 (1998).

[87] Jogl, G. & Tong, L. Crystal structure of yeast acetyl-coenzyme a synthetase in complex with amp. *Biochemistry* **43**, 1425–31 (2004).

[88] Nakatsu, T. *et al.* Structural basis for the spectral difference in luciferase bioluminescence. *Nature* **440**, 372–6 (2006).

[89] Hisanaga, Y. *et al.* Structural basis of the substrate-specific two-step catalysis of long chain fatty acyl-coa synthetase dimer. *J Biol Chem* **279**, 31717–26 (2004).

[90] Kochan, G., Pilka, E. S., von Delft, F., Oppermann, U. & Yue, W. W. Structural snapshots for the conformation-dependent catalysis by human medium-chain acyl-coenzyme a synthetase acsm2a. *Journal of Molecular Biology* **388**, 997 – 1008 (2009).

[91] Reger, A. S., Carney, J. M. & Gulick, A. M. Biochemical and crystallographic analysis of substrate binding and conformational changes in acetyl-coa synthetase. *Biochemistry* **46**, 6536–46 (2007).

[92] Starai, V. J., Celic, I., Cole, R. N., Boeke, J. D. & Escalante-Semerena, J. C. Sir2-dependent activation of acetyl-coa synthetase by deacetylation of active lysine. *Science* **298**, 2390–2 (2002).

[93] Starai, V. J., Takahashi, H., Boeke, J. D. & Escalante-Semerena, J. C. Short-chain fatty acid activation by acyl-coenzyme a synthetases requires sir2 protein function in salmonella enterica and saccharomyces cerevisiae. *Genetics* **163**, 545–55 (2003).

[94] Branchini, B. R., Murtiashaw, M. H., Magyar, R. A. & Anderson, S. M. The role of lysine 529, a conserved residue of the acyl-adenylate-forming enzyme superfamily, in firefly luciferase. *Biochemistry* **39**, 5433–40 (2000).

[95] Horswill, A. R. & Escalante-Semerena, J. C. Characterization of the propionyl-coa synthetase (prpe) enzyme of salmonella enterica: residue lys592 is required for propionyl-amp synthesis. *Biochemistry* **41**, 2379–87 (2002).

[96] Tomino, S., Yamada, M., Itoh, H. & Kurahashik. Cell-free synthesis of gramicidin s. *Biochemistry* **6**, 2552–60 (1967).

[97] Saito, M., Hori, K., Kurotsu, T., Kanda, M. & Saito, Y. Three conserved glycine

residues in valine activation of gramicidin s synthetase 2 from bacillus brevis. *J Biochem* **117**, 276–82 (1995).

[98] Fry, D. C., Kuby, S. A. & Mildvan, A. S. Atp-binding site of adenylate kinase: mechanistic implications of its homology with ras-encoded p21, f1-atpase, and other nucleotide-binding proteins. *Proc Natl Acad Sci U S A* **83**, 907–11 (1986).

[99] Pai, E. F. *et al.* Structure of the guanine-nucleotide-binding domain of the ha-ras oncogene product p21 in the triphosphate conformation. *Nature* **341**, 209–14 (1989).

[100] Gocht, M. & Marahiel, M. A. Analysis of core sequences in the d-phe activating domain of the multifunctional peptide synthetase tyca by site-directed mutagenesis. *J Bacteriol* **176**, 2654–62 (1994).

[101] Stuible, H., Buttner, D., Ehlting, J., Hahlbrock, K. & Kombrink, E. Mutational analysis of 4-coumarate:coa ligase identifies functionally important amino acids and verifies its close relationship to other adenylate-forming enzymes. *FEBS Lett* **467**, 117–22 (2000).

[102] Weckermann, R., Furbass, R. & Marahiel, M. A. Complete nucleotide sequence of the tyca gene coding the tyrocidine synthetase 1 from bacillus brevis. *Nucleic Acids Res* **16**, 11841 (1988).

[103] Lee, S. G., Roskoski, J., R., Bauer, K. & Lipmann, F. Purification of the polyenzymes responsible for tyrocidine synthesis and their dissociation into subunits. *Biochemistry* **12**, 398–405 (1973).

[104] Chang, K. H., Xiang, H. & Dunaway-Mariano, D. Acyl-adenylate motif of the acyl-adenylate/thioester-forming enzyme superfamily: a site-directed mutagenesis study with the pseudomonas sp. strain cbs3 4-chlorobenzoate:coenzyme a ligase. *Biochemistry* **36**, 15650–9 (1997).

[105] Pavela-Vrancic, M. *et al.* Atp binding in peptide synthetases: determination of contact sites of the adenine moiety by photoaffinity labeling of tyrocidine synthetase 1 with 2-azidoadenosine triphosphate. *Biochemistry* **33**, 6276–83 (1994).

[106] Serrano, R., Kielland-Brandt, M. C. & Fink, G. R. Yeast plasma membrane atpase is essential for growth and has homology with (na+ + k+), k+- and ca2+-atpases. *Nature* **319**, 689–93 (1986).

[107] Addison, R. Primary structure of the neurospora plasma membrane h+-atpase deduced from the gene sequence. homology to na+/k+-, ca2+-, and k+-atpase. *Journal of Biological Chemistry* **261**, 14896–901 (1986).

[108] Taylor, W. R. & Green, N. M. The predicted secondary structures of the nucleotide-binding sites of six cation-transporting atpases lead to a probable tertiary fold. *Eur J Biochem* **179**, 241–8 (1989).

[109] Serrano, R. Structure and function of proton translocating atpase in plasma membranes of plants and fungi. *Biochim Biophys Acta* **947**, 1–28 (1988).

[110] Inohara, N. *et al.* Two genes, atpc1 and atpc2, for the gamma subunit of arabidopsis thaliana chloroplast atp synthase. *J Biol Chem* **266**, 7333–8 (1991).

[111] Blanco, F. J., Rivas, G. & Serrano, L. A short linear peptide that folds into a native stable beta-hairpin in aqueous solution. *Nature Structural Biology* **1**, 584–90 (1994).

[112] Tokita, K., Hori, K., Kurotsu, T., Kanda, M. & Saito, Y. Effect of single base substitutions at glycine-870 codon of gramicidin s synthetase 2 gene on proline activation. *J Biochem* **114**, 522–7 (1993).

[113] Dieckmann, R., Pavela-Vrancic, M., Pfeifer, E., von Dohren, H. & Kleinkauf, H. The adenylation domain of tyrocidine synthetase 1–structural and functional role of the interdomain linker region and the (s/t)gt(t/s)gxpkg core sequence. *Eur J Biochem* **247**, 1074–82 (1997).

[114] Drake, E. J., Nicolai, D. A. & Gulick, A. M. Structure of the entb multidomain nonribosomal peptide synthetase and functional analysis of its interaction with the ente adenylation domain. *Chem Biol* **13**, 409–19 (2006).

[115] Wu, R. *et al.* Mechanism of 4-chlorobenzoate:coenzyme a ligase catalysis. *Biochemistry* **47**, 8026–39 (2008).

[116] Schlumbohm, W. *et al.* An active serine is involved in covalent substrate amino acid binding at each reaction center of gramicidin s synthetase. *J Biol Chem* **266**, 23135–41 (1991).

[117] De Crecy-Lagard, V., Marliere, P. & Saurin, W. Multienzymatic non ribosomal peptide biosynthesis: identification of the functional domains catalysing peptide elongation and epimerisation. *C R Acad Sci III* **318**, 927–36 (1995).

[118] Lambalot, R. H. *et al.* A new enzyme superfamily - the phosphopantetheinyl transferases. *Chem Biol* **3**, 923–36 (1996).

[119] Parris, K. D. *et al.* Crystal structures of substrate binding to bacillus subtilis holo-(acyl carrier protein) synthase reveal a novel trimeric arrangement of molecules resulting in three active sites. *Structure* **8**, 883–95 (2000).

[120] Flugel, R. S., Hwangbo, Y., Lambalot, R. H., Cronan, J., J. E. & Walsh, C. T. Holo-(acyl carrier protein) synthase and phosphopantetheinyl transfer in escherichia coli. *J Biol Chem* **275**, 959–68 (2000).

[121] Chirgadze, N. Y., Briggs, S. L., McAllister, K. A., Fischl, A. S. & Zhao, G. Crystal structure of streptococcus pneumoniae acyl carrier protein synthase: an essential enzyme in bacterial fatty acid biosynthesis. *Embo Journal* **19**, 5281–7 (2000).

[122] Mofid, M. R., Finking, R. & Marahiel, M. A. Recognition of hybrid peptidyl carrier proteins/acyl carrier proteins in nonribosomal peptide synthetase modules by the 4'-phosphopantetheinyl transferases acps and sfp. *J Biol Chem* **277**, 17023–31 (2002).

[123] Gehring, A. M., Lambalot, R. H., Vogel, K. W., Drueckhammer, D. G. & Walsh, C. T. Ability of streptomyces spp. acyl carrier proteins and coenzyme a analogs to serve as substrates in vitro for e. coli holo-acp synthase. *Chem Biol* **4**, 17–24 (1997).

[124] Quadri, L. E. *et al.* Characterization of sfp, a bacillus subtilis phosphopantetheinyl transferase for peptidyl carrier protein domains in peptide synthetases. *Biochemistry* **37**, 1585–95 (1998).

[125] Mootz, H. D., Finking, R. & Marahiel, M. A. 4'-phosphopantetheine transfer in primary and secondary metabolism of bacillus subtilis. *J Biol Chem* **276**, 37289–98 (2001).

[126] Pfeifer, E., Pavela-Vrancic, M., von Dohren, H. & Kleinkauf, H. Characterization of tyrocidine synthetase 1 (ty1): requirement of posttranslational modification for peptide biosynthesis. *Biochemistry* **34**, 7450–9 (1995).

[127] Stachelhaus, T., Huser, A. & Marahiel, M. A. Biochemical characterization of peptidyl carrier protein (pcp), the thiolation domain of multifunctional peptide synthetases. *Chem Biol* **3**, 913–21 (1996).

[128] Spivey, H. O. & Ovadi, J. Substrate channeling. *Methods* **19**, 306–21 (1999).

[129] Weissman, K. J. & Muller, R. Crystal structure of a molecular assembly line. *Angew Chem Int Ed Engl* 1521–3773 (2008).

[130] Perham, R. N. Swinging arms and swinging domains in multifunctional enzymes: catalytic machines for multistep reactions. *Annu Rev Biochem* **69**, 961–1004 (2000).

[131] Schwarzer, D., Mootz, H. D., Linne, U. & Marahiel, M. A. Regeneration of misprimed nonribosomal peptide synthetases by type ii thioesterases. *Proc Natl Acad Sci U S A* **99**, 14083–8 (2002).

[132] Weber, W., Hunenberger, P. H. & McCammon, J. A. Molecular dynamics simulations of a polyalanine octapeptide under ewald boundary conditions; influence of artificial periodicity on peptide conformation. *The journal of physical chemistry B* **104**, 3668–75 (2000).

[133] Kim, Y. & Prestegard, J. H. A dynamic model for the structure of acyl carrier protein in solution. *Biochemistry* **28**, 8792–7 (1989).

[134] Crump, M. P. *et al.* Solution structure of the actinorhodin polyketide synthase acyl carrier protein from streptomyces coelicolor a3(2). *Biochemistry* **36**, 6000–8 (1997).

[135] Holak, T. A., Kearsley, S. K., Kim, Y. & Prestegard, J. H. Three-dimensional structure of acyl carrier protein determined by nmr pseudoenergy and distance geometry calculations. *Biochemistry* **27**, 6135–42 (1988).

[136] Holak, T. A., Nilges, M., Prestegard, J. H., Gronenborn, A. M. & Clore, G. M. Three-dimensional structure of acyl carrier protein in solution determined by nuclear magnetic resonance and the combined use of dynamical simulated annealing and distance geometry. *Eur J Biochem* **175**, 9–15 (1988).

[137] Xu, G. Y. *et al.* Solution structure of b. subtilis acyl carrier protein. *Structure* **9**, 277–87 (2001).

[138] Wong, H. C., Liu, G., Zhang, Y. M., Rock, C. O. & Zheng, J. The solution structure of acyl carrier protein from mycobacterium tuberculosis. *J Biol Chem* **277**, 15874–80 (2002).

[139] Findlow, S. C., Winsor, C., Simpson, T. J., Crosby, J. & Crump, M. P. Solution structure and dynamics of oxytetracycline polyketide synthase acyl carrier protein from streptomyces rimosus. *Biochemistry* **42**, 8423–33 (2003).

[140] Li, Q., Khosla, C., Puglisi, J. D. & Liu, C. W. Solution structure and backbone dynamics of the holo form of the frenolicin acyl carrier protein. *Biochemistry* **42**, 4648–57 (2003).

[141] Koglin, A. *et al.* Conformational switches modulate protein interactions in peptide antibiotic synthetases. *Science* **312**, 273–6 (2006).

[142] Finking, R., Mofid, M. R. & Marahiel, M. A. Mutational analysis of peptidyl carrier protein and acyl carrier protein synthase unveils residues involved in protein-protein recognition. *Biochemistry* **43**, 8946–56 (2004).

[143] Lai, J. R., Fischbach, M. A., Liu, D. R. & Walsh, C. T. Localized protein interaction surfaces on the entb carrier protein revealed by combinatorial mutagenesis and selection. *J Am Chem Soc* **128**, 11002–3 (2006).

[144] Zhou, Z., Lai, J. R. & Walsh, C. T. Interdomain communication between the thiolation and thioesterase domains of entf explored by combinatorial mutagenesis and selection. *Chem Biol* **13**, 869–79 (2006).

[145] Lai, J. R., Fischbach, M. A., Liu, D. R. & Walsh, C. T. A protein interaction surface in nonribosomal peptide synthesis mapped by combinatorial mutagenesis and selection. *Proc Natl Acad Sci U S A* **103**, 5314–9 (2006).

[146] Zhou, Z., Lai, J. R. & Walsh, C. T. Directed evolution of aryl carrier proteins in the enterobactin synthetase. *Proc Natl Acad Sci U S A* **104**, 11621–6 (2007).

[147] Lipmann, F., Gevers, W., Kleinkauf, H. & Roskoski, J., R. Polypeptide synthesis on protein templates: the enzymatic synthesis of gramicidin s and tyrocidine. *Adv Enzymol Relat Areas Mol Biol* **35**, 1–34 (1971).

[148] Keating, T. A., Marshall, C. G., Walsh, C. T. & Keating, A. E. The structure of vibh represents nonribosomal peptide synthetase condensation, cyclization and epimerization domains. *Nat Struct Biol* **9**, 522–6 (2002).

[149] Leslie, A. G. Refined crystal structure of type iii chloramphenicol acetyltransferase at 1.75 a resolution. *J Mol Biol* **213**, 167–86 (1990).

[150] Leslie, A. G., Moody, P. C. & Shaw, W. V. Structure of chloramphenicol acetyltransferase at 1.75-a resolution. *Proc Natl Acad Sci U S A* **85**, 4133–7 (1988).

[151] Mattevi, A. *et al.* Atomic structure of the cubic core of the pyruvate dehydrogenase multienzyme complex. *Science* **255**, 1544–50 (1992).

[152] Mattevi, A., Obmolova, G., Kalk, K. H., Teplyakov, A. & Hol, W. G. Crystallographic analysis of substrate binding and catalysis in dihydrolipoyl transacetylase (e2p). *Biochemistry* **32**, 3887–901 (1993).

[153] Mattevi, A. *et al.* Refined crystal structure of the catalytic domain of dihydrolipoyl transacetylase (e2p) from azotobacter vinelandii at 2.6 a resolution. *J Mol Biol* **230**, 1183–99 (1993).

[154] Ehmann, D. E., Trauger, J. W., Stachelhaus, T. & Walsh, C. T. Aminoacyl-snacs as small-molecule substrates for the condensation domains of nonribosomal peptide synthetases. *Chem Biol* **7**, 765–72 (2000).

[155] Hoffmann, K., Schneider-Scherzer, E., Kleinkauf, H. & Zocher, R. Purification and characterization of eucaryotic alanine racemase acting as key enzyme in cyclosporin biosynthesis. *J Biol Chem* **269**, 12710–4 (1994).

[156] Stachelhaus, T. & Walsh, C. T. Mutational analysis of the epimerization domain in the initiation module pheate of gramicidin s synthetase. *Biochemistry* **39**, 5775–87 (2000).

[157] Patel, H. M., Tao, J. & Walsh, C. T. Epimerization of an l-cysteinyl to a d-cysteinyl residue during thiazoline ring formation in siderophore chain elongation by pyochelin synthetase from pseudomonas aeruginosa. *Biochemistry* **42**, 10514–27 (2003).

[158] Clugston, S. L., Sieber, S. A., Marahiel, M. A. & Walsh, C. T. Chirality of peptide bond-forming condensation domains in nonribosomal peptide synthetases: the c5 domain of tyrocidine synthetase is a (d)c(l) catalyst. *Biochemistry* **42**, 12095–104 (2003).

[159] Linne, U., Doekel, S. & Marahiel, M. A. Portability of epimerization domain and role of peptidyl carrier protein on epimerization activity in nonribosomal peptide synthetases. *Biochemistry* **40**, 15824–34 (2001).

[160] Kohli, R. M., Trauger, J. W., Schwarzer, D., Marahiel, M. A. & Walsh, C. T. Generality of peptide cyclization catalyzed by isolated thioesterase domains of nonribosomal peptide synthetases. *Biochemistry* **40**, 7099–108 (2001).

[161] Finking, R. & Marahiel, M. A. Biosynthesis of nonribosomal peptides1. *Annu Rev Microbiol* **58**, 453–88 (2004).

[162] Hoyer, K. M., Mahlert, C. & Marahiel, M. A. The iterative gramicidin s thioesterase catalyzes peptide ligation and cyclization. *Chem Biol* **14**, 13–22 (2007).

[163] Bruner, S. D. *et al.* Structural basis for the cyclization of the lipopeptide antibiotic surfactin by the thioesterase domain srfte. *Structure* **10**, 301–10 (2002).

[164] Tseng, C. C. *et al.* Characterization of the surfactin synthetase c-terminal thioesterase domain as a cyclic depsipeptide synthase. *Biochemistry* **41**, 13350–9 (2002).

[165] Reimmann, C. *et al.* Essential pchg-dependent reduction in pyochelin biosynthesis of pseudomonas aeruginosa. *J Bacteriol* **183**, 813–20 (2001).

[166] Du, L., Chen, M., Sanchez, C. & Shen, B. An oxidation domain in the blmiii non-ribosomal peptide synthetase probably catalyzing thiazole formation in the biosynthesis of the anti-tumor drug bleomycin in streptomyces verticillus atcc15003. *FEMS Microbiol Lett* **189**, 171–5 (2000).

[167] Silakowski, B. *et al.* New lessons for combinatorial biosynthesis from myxobacteria. the myxothiazol biosynthetic gene cluster of stigmatella aurantiaca dw4/3-1. *J Biol Chem* **274**, 37391–9 (1999).

[168] Schneider, T. L., Shen, B. & Walsh, C. T. Oxidase domains in epothilone and bleomycin biosynthesis: thiazoline to thiazole oxidation during chain elongation. *Biochemistry* **42**, 9722–30 (2003).

[169] Weinig, S., Mahmud, T. & Muller, R. Markerless mutations in the myxothiazol biosynthetic gene cluster: a delicate megasynthetase with a superfluous nonribosomal peptide synthetase domain. *Chem Biol* **10**, 953–60 (2003).

[170] Mitchell, C. A., Shi, C., Aldrich, C. C. & Gulick, A. M. Structure of pa1221, a nonribosomal peptide synthetase containing adenylation and peptidyl carrier protein domains. *Biochemistry* **51**, 3252–3263 (2012).

[171] Anfinsen, C. B., Haber, E., Sela, M. & White, F. W. R. The kinetics of the formation of native ribonuclease during oxidation of the reduced polypeptide domain. *Proceedings of the National Academy of Sciences of the United States of America* **47**, 1309–14 (1961).

[172] Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **181**, 223–230 (1973).

[173] van Gunsteren, W. Molecular dynamics studies of proteins. *Current Opinion in Structural Biology* **3**, 167–74 (1993).

[174] Chou, P. Y. & Fasman, G. D. Prediction of protein conformation. *Biochemistry* **13**, 222–45 (1974).

[175] Garnier, J., Osguthorpe, D. J. & Robson, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol* **120**, 97–120 (1978).

[176] Jones, D. T. Protein structure prediction. In Orengo, C. A., Jones, D. T. & Thornton, J. M. (eds.) *Bioinformatics. Genes, Proteins and Computers*, 135–150 (BIOS, Oxford, 2003).

[177] Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology* **292**, 195–202 (1999).

[178] Rost, B. & Sander, C. Jury returns on structure prediction. *Nature* **360**, 540 (1992).

[179] Rost, B. & Sander, C. Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology* 584–599 (1993).

[180] Rost, B. Twilight zone of protein sequence alignments. *Protein Eng* **12**, 85–94 (1999).

[181] Rost, B. Prediction in 1d: secondary structure, membrane helices, and accessibility. *Methods Biochem Anal* **44**, 559–87 (2003).

[182] Martin, A. C. R. Comparative modeling. In Orengo, C. A., Jones, D. T. & Thornton, J. M. (eds.) *Bioinformatics. Genes, Proteins and Computers*, Advanced Text, 121–133 (BIOS, Oxford, 2003).

[183] Pieper, U., Eswar, N., Stuart, A. C., Ilyin, V. A. & Sali, A. Modbase, a database of annotated comparative protein structure models. *Nucleic Acids Res* **30**, 255–9 (2002).

[184] Pieper, U. *et al.* Modbase: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res* **34**, D291–5 (2006).

[185] Krieger, E., Nabuurs, S. B. & Vriend, G. Homology modeling. In Bourne, P. E. & Weissig, H. (eds.) *Structural Bioinformatics*, 509–24 (John Wiley & Sons, New Jersey, 2003).

[186] Greer, J. Comparative model-building of the mammalian serine proteases. *J Mol Biol* **153**, 1027–42 (1981).

[187] Altschul, S. F. *et al.* Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–402 (1997).

[188] Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* **85**, 2444–8 (1988).

[189] Marti-Renom, M. A. *et al.* Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* **29**, 291–325 (2000).

[190] Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* **234**, 779–815 (1993).

[191] Wallner, B. & Elofsson, A. All are not equal: a benchmark of different homology modeling programmes. *Protein Science* **14**, 1315–1327 (2005).

[192] Sutcliffe, M. J., Haneef, I., Carney, D. & Blundell, T. L. Knowledge based modelling of homologous proteins, part i: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng* **1**, 377–84 (1987).

[193] MacKerell, A. D. *et al.* All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry B* **102**, 3586–3616 (1998).

[194] Fan, H. & Mark, A. E. Refinement of homology-based protein structures by molecular dynamics simulation techniques. *Protein Science* **13**, 211–20 (2004).

[195] Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J Mol Biol* **7**, 95–9 (1963). 0022-2836 (Print) Journal Article.

[196] Laskowski, R. A. Structural quality assurance. In Bourne, P. E. & Weissig, H. (eds.) *Structural Bioinformatics*, Advanced Text, 273–304 (BIOS, Oxford, 2003).

[197] Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. Procheck: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography* **26**, 283–291 (1993).

[198] Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. Stereochemical quality of protein structure coordinates. *Proteins* **12**, 345–64 (1992).

[199] Lesk, A. M. Alignments and phylogenetic trees. In *Introduction to Bioinformatics*, 157 – 210 (Oxford University Press, Oxford, 2005), 2nd edn.

[200] Sippl, M. J. Recognition of errors in three-dimensional structures of proteins. *Proteins* **17**, 355–62 (1993).

[201] Shen, M. Y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci* **15**, 2507–24 (2006).

[202] Warren, G. L. *et al.* A critical assessment of docking programs and scoring functions. *Journal of Medicinal Chemistry* **49**, 5912–31 (2006).

[203] Leach, A. R. Computer simulation methods. In *Molecular Modelling Principles and Applications*, 312–323 (Pearson, Harrow, 2001), 2nd edn.

[204] Kuntz, I. D. A geometric approach to macromolecule-ligand interactions. *Journal of molecular biology* **161**, 269–288 (1982).

[205] C, B. R., Subramanian, J. & Sharma, S. D. Managing protein flexibility in docking and its applications. *Drug Discov Today* **14**, 394–400 (2009).

[206] Sousa, S. F., Fernandes, P. A. & Ramos, M. J. Protein-ligand docking: current status and future challenges. *Proteins* **65**, 15–26 (2006).

[207] B′ohm, H. J. The computer program ludi: a new method for the de novo design of enzyme inhibitors. *Journal of computer-aided molecular design* **6**, 61–78 (1992).

[208] Mizutani, M. Y., Tomioka, N. & Itai, A. Rational automatic search method for stable docking models of protein and ligand. *J Mol Biol* **243**, 310–26 (1994).

[209] Rarey, M., Kramer, B., Lengauer, T. & Klebe, G. A fast flexible docking method using an incremental construction algorithm. *Journal of molecular biology* **261**, 470–89 (1996).

[210] Ewing, T. J. A. & Kuntz, I. D. Critical evaluation of search algorithms for automated molecular docking and database screening. *Journal of computational chemistry* **18**, 1175–1189 (1998).

[211] Miller, M. D. Flog: a system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure. *Journal of computer aided molecular design* **8**, 153–74 (1994).

[212] Hart, T. N. & Read, R. J. A multiple-start monte carlo docking method. *Proteins: Structure, Function, and Bioinformatics* **13**, 206–222 (1992).

[213] Hart, T. N., Ness, S. R. & Read, R. J. Critical evaluation of the research docking program for the casp2 challenge. *Proteins* **Suppl 1**, 205–9 (1997).

[214] Trosset, J. Y. & Scheraga, H. A. Prodock: Software package for protein modeling and docking. *Journal of Computational Chemistry* **20**, 412–427 (1999).

[215] Liu, M. & Wang, S. Mcdock: a monte carlo simulation approach to the molecular docking problem. *J Comput Aided Mol Des* **13**, 435–51 (1999).

[216] Holland, J. H. *Adaptation in Natural and Artificial systems* (University of Michigan Press, 1975).

[217] Jones, G., Willett, P. & Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J Mol Biol* **245**, 43–53 (1995).

[218] Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* **267**, 727–48 (1997).

[219] Clark, K. P. Flexible ligand docking without parameter adjustment across four ligand-receptor complexes. *Journal of Computational Chemistry* **16**, 1210–26 (1995).

[220] Morris, G. M. *et al.* Automated docking using a lamarckian genetic algorithm and and empirical binding free energy function. *J Comput Chem* **19**, 1639–62 (1998).

[221] Taylor, J. S. & Burnett, R. M. Darwin: a program for docking flexible molecules. *Proteins* **41**, 173–91 (2000).

[222] Morris, G. M. Autodock users guide (2001).

[223] Widom, B. The boltzmann distribution law and statistical thermodynamics. In *Statistical Mechanics A concise introduction for chemists*, 1–15 (Cambridge University Press, Cambridge, 2002).

[224] Diu, B., Guthmann, C., Lederer, D. & Roulet, B. About the fundamental postulate of statistical mechanics. *European Journal of Physics* **11**, 160–162 (1990).

[225] Cramer, C. J. Simulations of molecular ensembles. In *Essentials of computational chemistry Theories and Models*, 69–102 (John Wiley and Sons, Chicester, 2004), 2nd edn.

[226] Allen, M. P. & Tildesley, D. J. Statistical mechanics. In *Computer Simulation of Liquids*, 33 – 70 (Oxford Science Publications, Oxford, 2005).

[227] Hockney, R. W. The potential calculation and some applications. *Methods in Computational Physics* **9**, 136 – 211 (1970).

[228] van der Spoel, D. *et al.* *GROMACS User Manual Version 3.2*. www.gromacs.org (2004).

[229] Ryckaert, J. P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *Journal of Computational Physics* **23**, 327–41 (1977).

[230] Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. Lincs: A linear constraint solver for molecular simulations. *J. Comp. Chem.* **18**, 14631472 (1997).

[231] Moraitakis, G., Purkiss, A. G. & Goodfellow, J. M. Simualated dynamics and biological macromolecules. *Reports on Progress in Physics* **66**, 383 – 406 (2003).

[232] Born, M. & Von Karman, T. Uber chwingungen in raumgittern. *Physik Z* **13**, 297–309 (1912).

[233] van Gunsteren, W. F. *et al.* (eds.) *Biomolecular simulation:The GROMOS96 manual and user guide* (Zurich, Switzerland, 1996).

[234] Berendsen, H. J. C., van der Spoel, D. & van Drunen, R. Gromacs: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications* **91**, 43–56 (1995).

[235] Lindahl, E., Hess, B. & van der Spoel, D. Gromacs 3.0: A package for molecular simulation and trajectory analysis. *Journal of Molecular Modelling* **7**, 306–17 (2001).

[236] MacKerell, J., A. D. Empirical force fields for biological macromolecules. *Journal of Computational Chemistry* **25**, 1584–1604 (2004).

[237] Cramer, C. J. Molecular mechanics. In *Essentials of computational chemistry Theories and Models*, 17–67 (John Wiley and Sons, Chicester, 2004), 2nd edn.

[238] Darden, T., York, D. & Pedersen, L. Particle mesh ewald: An n-log(n) method for ewald sums in large systems. *J. Chem. Phys.* **98**, 1008910092 (1993).

[239] Essmann, U. *et al.* A smooth particle mesh ewald potential. *Journal of Chemical Physics* **103**, 8577–8592 (1995).

[240] Ewald, P. P. Die berechnung optischer und elektrostatischer gitterpotential. *Ann Phy* **64**, 253 – 287 (1921).

[241] Zielkiewicz, J. Structural properties of water: Comparison of the spc, spce, tip4p, and tip5p models of water. *The Journal of chemical physics* **123**, 104501.1–104501.6 (2005).

[242] Berendsen, H. J. C., Postma, J. P. M., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *Journal of Chemical Physics* **81**, 3684–3690 (1984).

[243] Nose, S. A molecular dynamics method for simulations in the canonical ensemble. *Molecular Physics* **50**, 255–268 (1984).

[244] Hoover, W. G. Canonical dynamics: equilibrium phase-space distributions. *Physical Review A* **31**, 1695 – 1697 (1985).

[245] Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **52**, 7182 – 7190 (1981).

[246] McCammon, J. A., Gelin, B. R. & Karplus, M. Dynamics of folded proteins. *Nature* **267**, 585–90 (1977).

[247] Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat Struct Biol* **9**, 646–52 (2002).

[248] Hansson, T., Oostenbrink, C. & van Gunsteren, W. Molecular dynamics simulations. *Curr Opin Struct Biol* **12**, 190–6 (2002).

[249] MacKerell, A. D., Jr. *et al.* Self-consistent parameterization of biomolecules for molecular modeling and condensed phase simulations. *The Journal of the Federation of American Societies for Experimental Biology* **6**, A143 (1992).

[250] Phillips, J. C. *et al.* Scalable molecular dynamics with namd. *J Comput Chem* **26**, 1781–802 (2005).

[251] Freddolino, P. L., Arkhipov, A. S., Larson, S. B., McPherson, A. & Schulten, K. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure* **14**, 437–49 (2006).

[252] Day, J., Kuznetsov, Y. G., Larson, S. B., Greenwood, A. & McPherson, A. Biophysical studies on the rna cores of satellite tobacco mosaic virus. *Biophys J* **80**, 2364–71 (2001).

[253] Larson, S. B. & McPherson, A. Satellite tobacco mosaic virus rna: structure and implications for assembly. *Curr Opin Struct Biol* **11**, 59–65 (2001).

[254] Kuznetsov, Y. G., Daijogo, S., Zhou, J., Semler, B. L. & McPherson, A. Atomic force microscopy analysis of icosahedral virus rna. *J Mol Biol* **347**, 41–52 (2005).

[255] Jayachandran, G., Vishal, V. & Pande, V. S. Using massively parallel simulation and markovian models to study protein folding: examining the dynamics of the villin headpiece. *J Chem Phys* **124**, 164902 (2006).

[256] Shirts, M. & Pande, V. S. Computing: Screen savers of the world unite! *Science* **290**, 1903–1904 (2000).

[257] Rhee, Y. M., Sorin, E. J., Jayachandran, G., Lindahl, E. & Pande, V. S. Simulations of the role of water in the protein-folding mechanism. *Proc Natl Acad Sci U S A* **101**, 6456–61 (2004).

[258] Cornell, W. D. *et al.* A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society* **117**, 51795197 (1995).

[259] Garcia, A. E. & Sanbonmatsu, K. Y. Alpha-helical stabilization by side chain shielding of backbone hydrogen bonds. *Proc Natl Acad Sci U S A* **99**, 2782–7 (2002).

[260] van der Kamp, M. W., Shaw, K. E., Woods, C. J. & Mulholland, A. J. Biomolecular simulation and modelling: status, progress and prospects. *Journal of the Royal Society Interface* **5**, 173–190 (2008).

[261] Im, W. & Brooks, r., C. L. Interfacial folding and membrane insertion of designed peptides studied by molecular dynamics simulations. *Proc Natl Acad Sci U S A* **102**, 6771–6 (2005).

[262] Levitt, M. & Warshel, A. Computer simulation of protein folding. *Nature* **253**, 694–8 (1975).

[263] Shelley, J. C., Shelley, M. Y., Reeder, R. C., Bandyopadhyay, S. & Klein, M. L. A coarse grain model for phospholipid simulations. *Journal of Physical Chemistry B* **105**, 4464–4470 (2001).

[264] Arkhipov, A., Freddolino, P. L. & Schulten, K. Stability and dynamics of virus capsids described by coarse-grained modeling. *Structure* **14**, 1767–77 (2006).

[265] Neri, M., Anselmi, C., Cascella, M., Maritan, A. & Carloni, P. Coarse-grained model of proteins incorporating atomistic detail of the active site. *Phys Rev Lett* **95**, 218102 (2005).

[266] Guex, N. & Peitsch, M. C. Swiss-model and the swiss-pdb viewer: An environment for comparative protein modeling. *ELECTROPHORESIS* **18**, 2714–2723 (1997).

[267] Fiser, A., Do, R. K. & Sali, A. Modeling of loops in protein structures. *Protein Sci* **9**, 1753–73 (2000).

[268] van der Spoel, D., Feenstra, K. A., Hemminga, M. A. & Berendsen, H. J. Molecular modeling of the rna binding n-terminal part of cowpea chlorotic mottle virus coat protein in solution with phosphate ions. *Biophys J* **71** (1996).

[269] Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F. & Hermans, J. Interaction models for water in relation to protein hydration. In Pullman, B. (ed.) *Intermolecular Forces*, 331342 (D. Reidel Publishing Company, Dordrecht, 1981).

[270] Schuttelkopf, A. W. & van Aalten, D. M. Prodrg: a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallogr D Biol Crystallogr* **60**, 1355–63 (2004).

[271] van den Berg, P. A. W., Feenstra, K. A., Mark, A. E., Berendsen, H. J. C. & Visser, A. J. W. G. Dynamic conformations of flavin adenine dinucleotide: Simulated molecular dynamics of the flavin cofactor related to the time-resolved fluorescence characteristics. *Journal of Physical Chemistry B* **106**, 8858–69 (2002).

[272] Kosinsky, Y. A. *et al.* Development of the force field parameters for phosphoimidazole and phosphohistidine. *J Comput Chem* **25**, 1313–21 (2004).

[273] Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–637 (1983).

[274] Hayward, S. & Berendsen, H. J. Systematic analysis of domain motions in proteins from conformational change: new results on citrate synthase and t4 lysozyme. *Proteins* **30**, 144–54 (1998).

[275] Hayward, S. & Lee, R. A. Improvements in the analysis of domain motions in proteins from conformational change: Dyndom version 1.50. *J Mol Graph Model* **21**, 181–3 (2002).

[276] Pearson, K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2**, 559–572 (1901).

[277] Amadei, A., Linssen, B. M. A. & Berendsen, H. J. C. Essential dynamics of proteins. *Proteins: Structure, Function, and Bioinformatics* **17**, 412–425 (1993).

[278] Hayward, S. & Go, N. Collective variable description of native protein dynamics. *Annual Review of Physical Chemistry* **46**, 223–250 (1995).

[279] Di Vincenzo, L., Grgurina, I. & Pascarella, S. In silico analysis of the adenylation domains of the freestanding enzymes belonging to the eucaryotic nonribosomal peptide synthetase-like family. *Febs Journal* **272**, 929–41 (2005).

[280] Humphrey, W., Dalke, A. & Schulten, K. Vmd: visual molecular dynamics. *J Mol Graph* **14**, 33–8, 27–8 (1996).

[281] Demain, A. L. Pharmaceutically active secondary metabolites of microorganisms. *Appl Microbiol Biotechnol* **52**, 455–63 (1999).

[282] Hopwood, D. A. Forty years of genetics with streptomyces: from in vivo through in vitro to in silico. *Microbiology* **145 (Pt 9)**, 2183–202 (1999).

[283] Bentley, S. D. *et al.* Complete genome sequence of the model actinomycete streptomyces coelicolor a3(2). *Nature* **417**, 141–7 (2002).

[284] Challis, G. L. & Ravel, J. Coelichelin, a new peptide siderophore encoded by the streptomyces coelicolor genome: structure prediction from the sequence of its non-ribosomal peptide synthetase. *FEMS Microbiol Lett* **187**, 111–4 (2000).

[285] Challis, G. L., Ravel, J. & Townsend, C. A. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem Biol* **7**, 211–24 (2000).

[286] Redenbach, M. *et al.* A set of ordered cosmids and a detailed genetic and physical map for the 8 mb streptomyces coelicolor a3(2) chromosome. *Mol Microbiol* **21**, 77–96 (1996).

[287] Berman, H. M. *et al.* The protein data bank. *Nucleic Acids Res* **28**, 235–42 (2000).

[288] Rost, B. Marrying structure and genomics. *Structure* **6**, 259–63 (1998).

[289] Sanchez, R. & Sali, A. Large-scale protein structure modeling of the saccharomyces cerevisiae genome. *Proc Natl Acad Sci U S A* **95**, 13597–602 (1998).

[290] Bateman, A. *et al.* The pfam protein families database. *Nucleic Acids Res* **32**, D138–41 (2004).

[291] Notredame, C., Higgins, D. G. & Heringa, J. T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**, 205–17 (2000).

[292] Edgar, R. C. Muscle: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).

[293] Fischer, D. *et al.* Cafasp-1: critical assessment of fully automated structure prediction methods. *Proteins* **Suppl 3**, 209–17 (1999).

[294] Kelley, L. A., MacCallum, R. M. & Sternberg, M. J. Enhanced genome annotation using structural profiles in the program 3d-pssm. *J Mol Biol* **299**, 499–520 (2000).

[295] Bennett-Lovsey, R. M., Herbert, A. D., Sternberg, M. J. E. & Kelley, L. A. Exploring the extremes of sequence/structure space with ensemble fold recognition in

the program phyre. *Proteins: Structure, Function, and Bioinformatics* **70**, 611–625 (2008).

[296] Jones, D. T. Genthreader: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol* **287**, 797–815 (1999).

[297] McGuffin, L. J. & Jones, D. T. Improvement of the genthreader method for genomic fold recognition. *Bioinformatics* **19**, 874–81 (2003).

[298] Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292**, 195–202 (1999).

[299] Cuff, J. A. & Barton, G. J. Application of enhanced multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins: Structure, Function and Genetics* **40**, 502–511 (2000).

[300] Pollastri, G., Przybylski, D., Rost, B. & Baldi, P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* **47**, 228–35 (2002).

[301] Essmann, U. *et al.* A smooth particle mesh ewald potential. *J. Chem. Phys.* **103**, 85778592 (1995).

[302] Guex, N. & Peitsch, M. C. Swiss-pdbviewer: A fast and easy-to-use pdb viewer for macintosh and pc. *Protein Data Bank Quarterly Newsletter 77* **7** (1996).

[303] Barducci, A., Bonomi, M. & Parrinello, M. Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **1**, 826–843 (2011).

[304] Laio, A. & Parrinello, M. Escaping free-energy minima **99**, 12562–12566 (2002).

[305] Fischbach, M. A. & Walsh, C. T. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. *Chem Rev* **106**, 3468–96 (2006).

[306] Reuter, K., Mofid, M. R., Marahiel, M. A. & Ficner, R. Crystal structure of the surfactin synthetase-activating enzyme sfp: a prototype of the 4'-phosphopantetheinyl transferase superfamily. *Embo J* **18**, 6823–31 (1999).

[307] Richter, C. D., Nietlispach, D., Broadhurst, R. W. & Weissman, K. J. Multienzyme docking in hybrid megasynthetases. *Nat Chem Biol* **4**, 75–81 (2008).

[308] Sattely, E. S., Fischbach, M. A. & Walsh, C. T. Total biosynthesis: in vitro reconstitution of polyketide and nonribosomal peptide pathways. *Nat Prod Rep* **25**, 757–93 (2008).

[309] Walsh, C. T. *et al.* Tailoring enzymes that modify nonribosomal peptides during and after chain elongation on nrps assembly lines. *Curr Opin Chem Biol* **5**, 525–34 (2001).

[310] Keating, T. A. *et al.* Chain termination steps in nonribosomal peptide synthetase assembly lines: directed acyl-s-enzyme breakdown in antibiotic and siderophore biosynthesis. *Chembiochem* **2**, 99–107 (2001).